Data Science: Regression Models Course Project

Jagannatha Reddy

September 03, 2016

Problem Description

In this project we will be looking at a data set of a collection of cars in the **mtcars** available as part of the **datasets** package of R, and explore the relationship between a set of variables and miles per gallon (MPG) (outcome). Our main focus would be in answering the following two questions:

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

Executive Summary

Based on the analysis done on **mtcars** (from Motor Trend Car Road Tests data) data it is evident that Manual transmission vehicles give better MPG (Miles per US Gallon) compared to Automatic transmission vehicles. Also looking at the different models and exploration of data it is clear that other factors like number of Cylinders, and weight of the vehicle affect the MPG of the vehicle. This analysis is done only based on the dataset available only from 32 vehicles I strongly feel this data is not statistically significant. One more factor that should be considered is that data is obtained in 1974 and hence might be obsolete given the latest developments in the automotive industry

These conclusions are based on the limited dataset of 32. Though the results are consistent the data used for analysis doesn't appear to be statistically significant

1) Load and prepare the mtcars data

```
cache=TRUE
# load the required libraries
library(datasets)
data(mtcars)
#head(mtcars, 2)
mtcars$am <- as.factor(mtcars$am) # make the am value as factor
levels(mtcars$am) <- c("Automatic", "Manual") # 0 - Automatic, 1 - Manual
```

You can observe that there are totally 32 records in the mtcars dataset with each record having 11 columns to represent various properties (like MPG, Cylinders, Weight, etc.) of each of the vehicles

2) Is an automatic or manual transmission better for MPG

```
aggregate(mpg ~ am, data=mtcars, FUN=mean)
```

```
##          am      mpg
## 1 Automatic 17.14737
## 2    Manual 24.39231
```

You can observe that MPG value for Manual transmission is 24.39 which is significantly higher compared to that of Automatic transmission value of 17.15. Even the statistical inference & boxplot available in Appendix also concludes the same result

3) Quantify the MPG difference between automatic and manual transmissions

Let us explore how different vehicle properties influence the fuel efficiency of theAutomatic and Manual transmission vehicles

```
aModel1 <- lm(mpg ~ cyl, data=mtcars[mtcars$am=='Automatic',])
aModel2 <- lm(mpg ~ wt, data = mtcars[mtcars$am=='Automatic',])
aModel3 <- lm(mpg ~ cyl + wt, data = mtcars[mtcars$am=='Automatic',])
aModel4 <- lm(mpg ~ disp, data = mtcars[mtcars$am=='Automatic',])
aModel5 <- lm(mpg ~ disp + cyl, data = mtcars[mtcars$am=='Automatic',])
aModel6 <- lm(mpg ~ disp + cyl + wt, data = mtcars[mtcars$am=='Automatic',])
aModel7 <- lm(mpg ~ gear, data = mtcars[mtcars$am=='Automatic',])
```

These models result in the following R squared values for Automatic transmission vehicles

1. model-1: cylinders to mpg: **0.63**
2. model-2: weight to mpg: **0.59**
3. model-3: cylinders and weight to mpg: **0.76**
4. model-4: displacement to mpg: **0.63**
5. model-5: displacement and cylinders to mpg: **0.69**
6. model-6: displacement, cylinders and weight to mpg: **0.76**
7. model-7: gear to mpg: **0.29**

```
mModel1 <- lm(mpg ~ cyl, data=mtcars[mtcars$am=='Manual',])
mModel2 <- lm(mpg ~ wt, data = mtcars[mtcars$am=='Manual',])
mModel3 <- lm(mpg ~ cyl + wt, data = mtcars[mtcars$am=='Manual',])
mModel4 <- lm(mpg ~ disp, data = mtcars[mtcars$am=='Manual',])
mModel5 <- lm(mpg ~ disp + cyl, data = mtcars[mtcars$am=='Manual',])
mModel6 <- lm(mpg ~ disp + cyl + wt, data = mtcars[mtcars$am=='Manual',])
mModel7 <- lm(mpg ~ gear, data = mtcars[mtcars$am=='Manual',])
```

These models result in the following R squared values for Manual transmission vehicles

1. model-1: cylinders to mpg: **0.68**
2. model-2: weight to mpg: **0.83**
3. model-3: cylinders and weight to mpg: **0.84**
4. model-4: displacement to mpg: **0.7**
5. model-5: displacement and cylinders to mpg: **0.71**
6. model-6: displacement, cylinders and weight to mpg: **0.85**
7. model-7: gear to mpg: **0.16**

You can observe that R squared values are consistently high for cylinder & weight to mpg and displacement, cylinder, & weight to mpg for both categories of transmissions. This indicate these 2 combinations are better fits compared to the rest of the models.

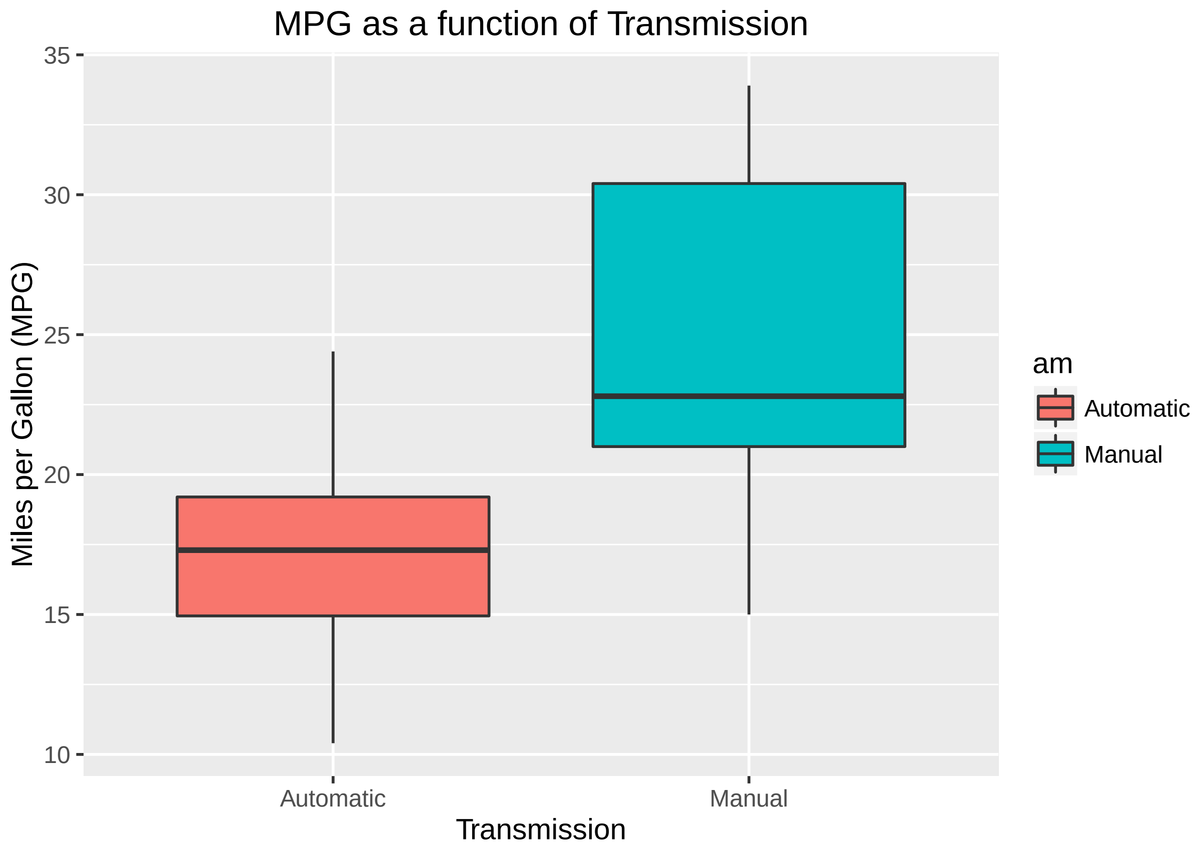
Based on the analysis done and also looking at the graphs in Apendix, you can observe that R squared values are consistently high for cylinder & weight to mpg and displacement, cylinder, & weight to mpg for both categories of transmissions. This indicate these 2 combinations are better fits compared to the rest of the models. Various models also indicate that fuel efficiency is influenced by several properties of the vehicles.

As stated eaerlier these conclusions are based on the limited dataset of 32 which isn't statistically significant.

APPENDIX

A) Boxplot of mpg as a function of transmission

```r
#boxplot of mpg as a function of transmission
library(ggplot2)
g<-ggplot(aes(y=mpg,x=am), data=mtcars)+geom_boxplot(aes(fill=am))
g+labs(x="Transmission", y="Miles per Gallon (MPG)", title="MPG as a function of Transmission")
```



B) Statistical Inference

Here we perform a t-test on mpg versus different transmissions

```r
#Perform Student's t-test on transmissions versus mpg
t.test(mpg~am, data=mtcars)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
```

```
## mean in group Automatic    mean in group Manual 
##                17.14737                24.39231
```

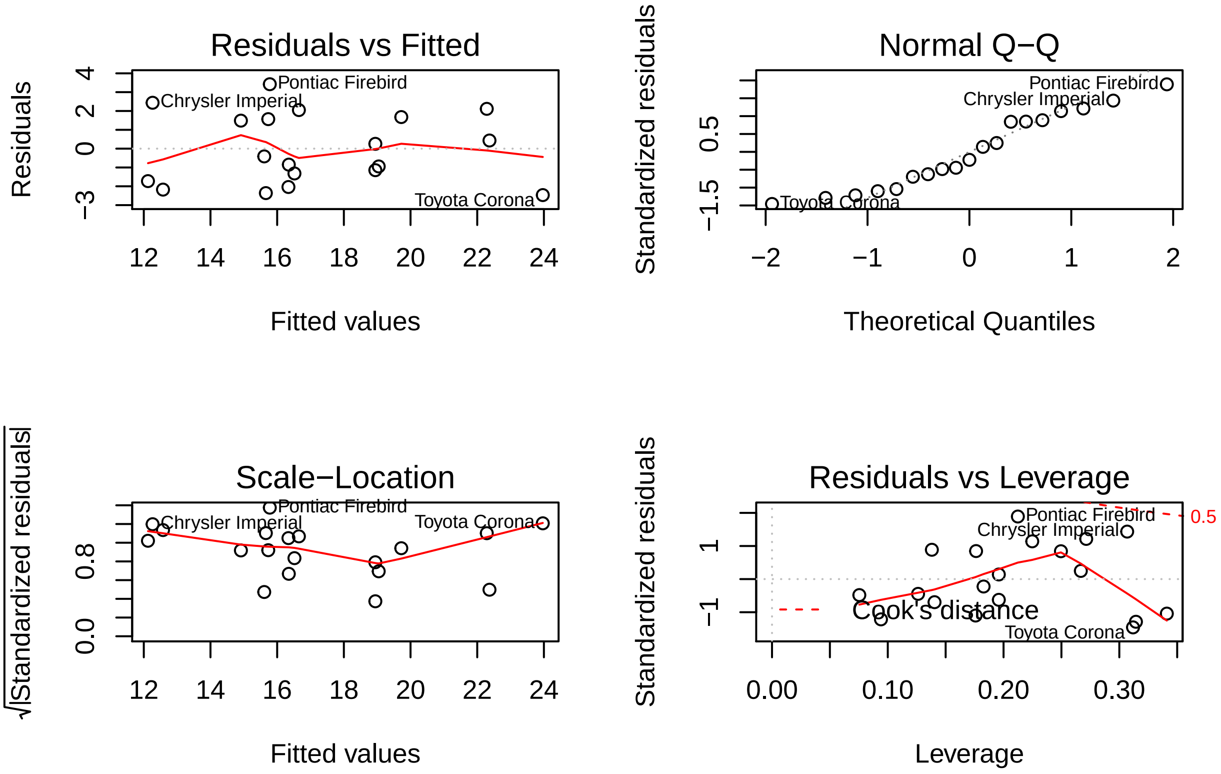
Based on the results, we can reject a null hypothesis that the effect of manual and automatic transmissions on mpg are the same

C) Residual Plot for displacement, cylinders and weight to mpg model

```
#boxplot of mpg as a function of transmission
summary(aModel6)
```

```
## 
## Call:
## lm(formula = mpg ~ disp + cyl + wt, data = mtcars[mtcars$am == 
##     "Automatic", ])
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -2.4668 -1.5218 -0.4123  1.6221  3.4210 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 35.188714   3.692450   9.530 9.38e-08 ***
## disp         0.002604   0.011170   0.233   0.8188    
## cyl         -1.396746   0.573364  -2.436   0.0278 *  
## wt          -2.412871   1.109103  -2.176   0.0460 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.039 on 15 degrees of freedom
## Multiple R-squared:  0.7642, Adjusted R-squared:  0.717 
## F-statistic:  16.2 on 3 and 15 DF,  p-value: 5.679e-05
```

```
par(mfrow=c(2,2))
plot(aModel6)
```



You can observe that R squared values are consistently high displacement, cylinder, & weight to mpg for both categories of transmissions. This indicate these 2 combinations are better fits compared to the rest of the models.