Data Science: Regression Models Course Project

Jagannatha Reddy

August 13, 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)
```

```
##               mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4      21   6  160 110  3.9 2.620 16.46  0  1    4    4
## Mazda RX4 Wag  21   6  160 110  3.9 2.875 17.02  0  1    4    4
```

```
mtcars$am <- as.factor(mtcars$am) # make the am value as factor
levels(mtcars$am) <- c("Automatic", "Manual") # 0 - Automatic, 1 - Manual
```

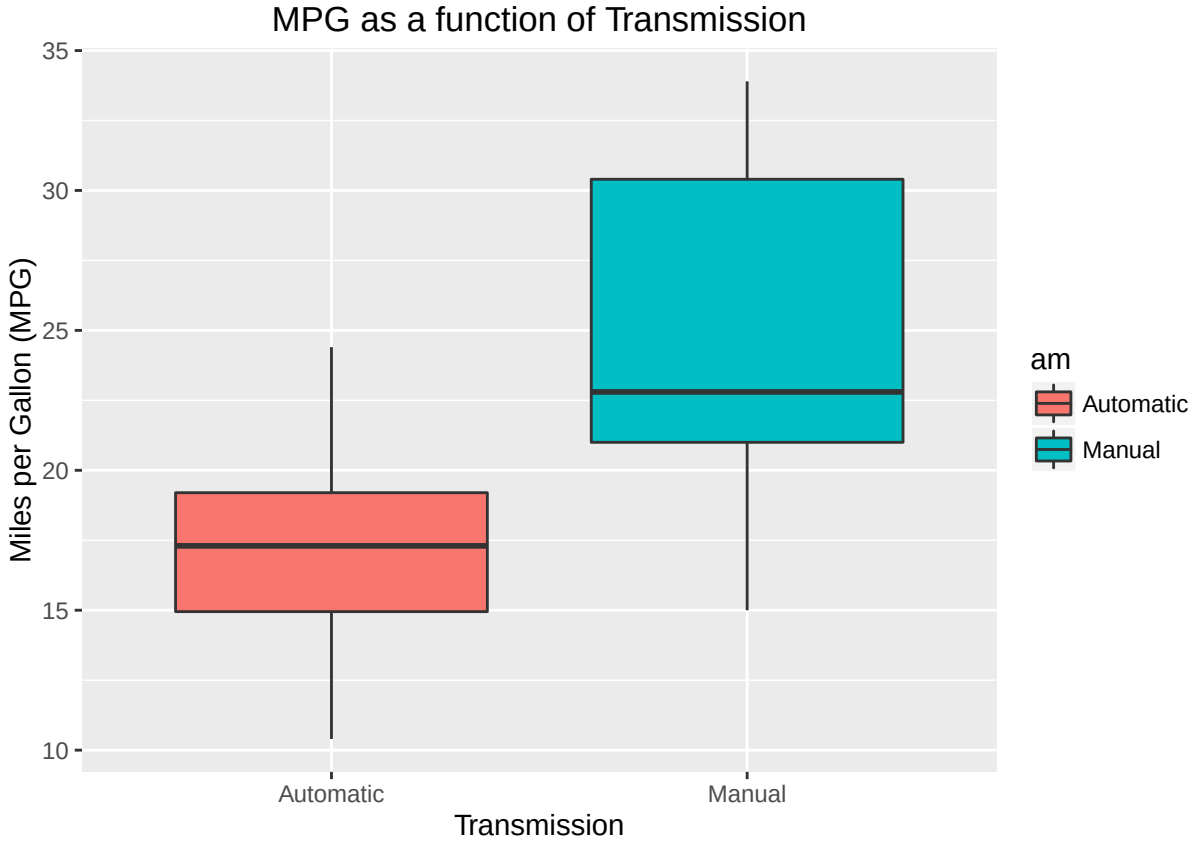
You can observe that there are totally 32 records in the mtcars dataset with each record having 11 columns to represent various properties (like MPG, Cylinders, Weight, etc.) of each of the vehicles

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
```

```
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")
```



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

3) Quantify the MPG difference between automatic and manual transmissions

Based on the analysis done in the **Different Models** section of 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

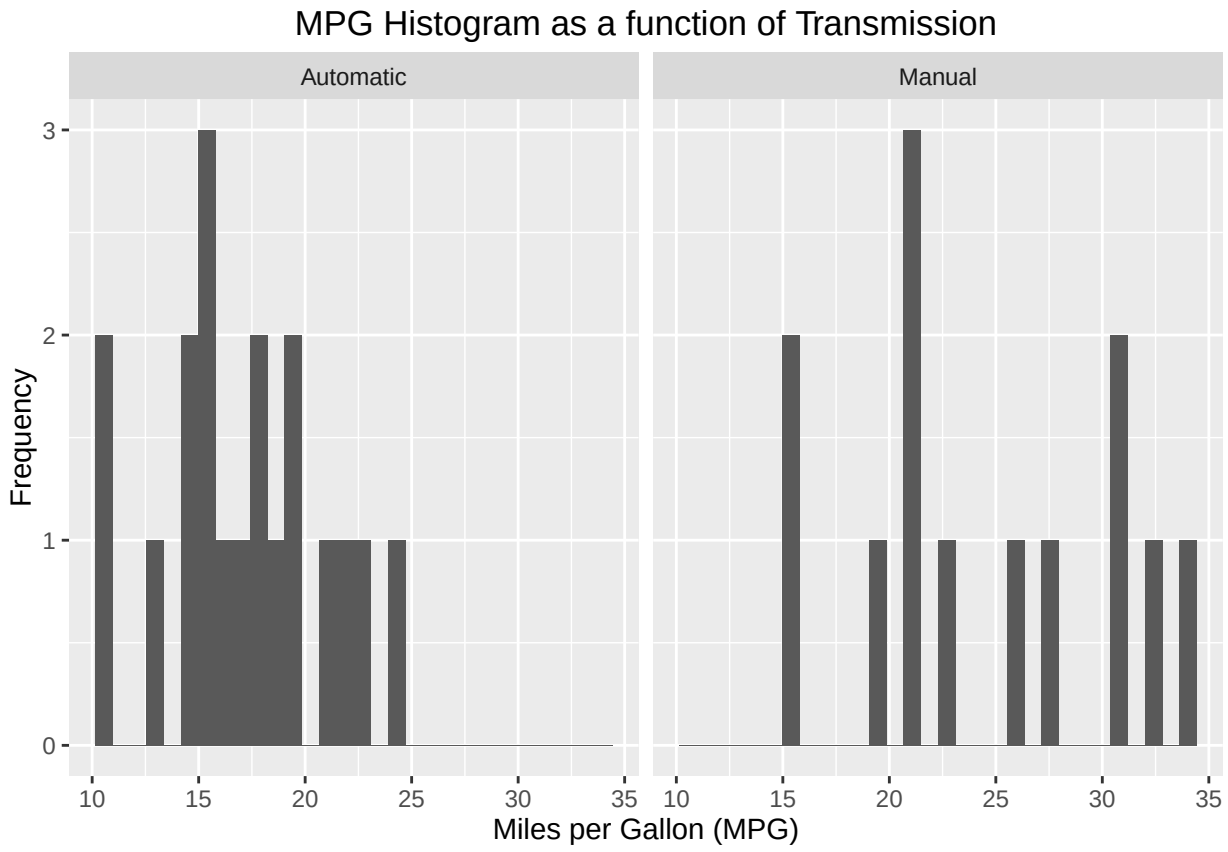
1) Exploratory Data Analysis

```
#convert to factors
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$gear <- as.factor(mtcars$gear)
mtcars$carb <- as.factor(mtcars$carb)
summary(mtcars)
```

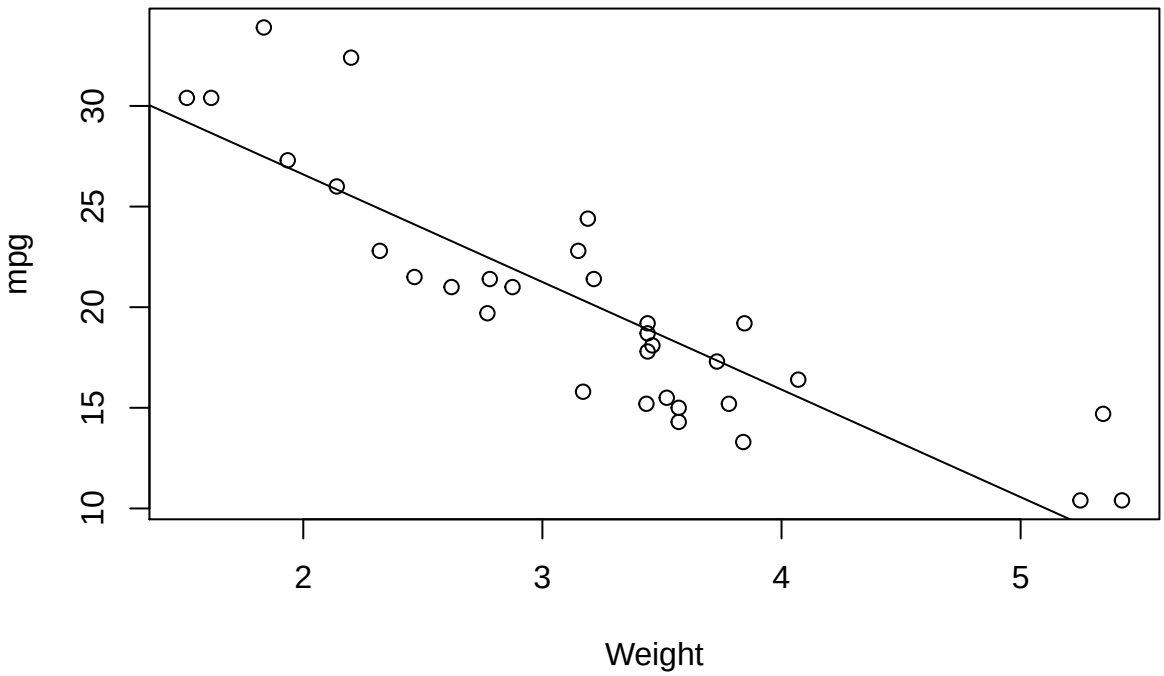
```
##       mpg        cyl         disp             hp             drat
##  Min.   :10.40   4:11   Min.   : 71.1   Min.   : 52.0   Min.   :2.760
##  1st Qu.:15.43   6: 7   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080
##  Median :19.20   8:14   Median :196.3   Median :123.0   Median :3.695
##  Mean   :20.09          Mean   :230.7   Mean   :146.7   Mean   :3.597
##  3rd Qu.:22.80          3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920
##  Max.   :33.90          Max.   :472.0   Max.   :335.0   Max.   :4.930
##        wt             qsec       vs              am     gear   carb
##  Min.   :1.513   Min.   :14.50   0:18   Automatic:19   3:15   1: 7
##  1st Qu.:2.581   1st Qu.:16.89   1:14   Manual   :13   4:12   2:10
##  Median :3.325   Median :17.71                         5: 5   3: 3
##  Mean   :3.217   Mean   :17.85                                4:10
##  3rd Qu.:3.610   3rd Qu.:18.90                                6: 1
##  Max.   :5.424   Max.   :22.90                                8: 1
```

```
g<-ggplot(data = mtcars, aes(mpg)) + geom_histogram() + facet_grid(.~am)
g+labs(x = "Miles per Gallon (MPG)", y = "Frequency", title = "MPG Histogram as a function of Transmissi
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#explore the relationship between weight and mpg
plot(mtcars$wt, mtcars$mpg, xlab="Weight", ylab="mpg")
fit<-lm(mpg~wt, data=mtcars)
abline(fit)
```



2) Different Models

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: **0.69**
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: **0.82**
6. model-6: displacement, cylinders and weight to mpg: **0.89**
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.