Motor Trend Car Road Tests - Effects of transmission on MPG

Nikhil

October 1, 2016

1. Executive Summary

This detailed analysis will be performed to fulfill the requirements of the course project for the course Regression Models offered by the Johns Hopkins University on Coursera. In this project, we will analyze the mtcars data set and explore the relationship between a set of variables and miles per gallon (MPG) which will be our outcome.

The main objectives of this research are as follows

- Is an automatic or manual transmission better for MPG?
- Quantifying how different is the MPG between automatic and manual transmissions?

The key takeway from our analysis was

- Manual transmission is better for MPG by a factor of 1.8 compared to automatic transmission.
- Means and medians for automatic and manual transmission cars are significantly different.

R version used for analysis is "R version 3.3.1 (2016-06-21)". OS used is "x86_64, Windows 10 64 Bit".

2. Basic settings

```r
echo = TRUE  # Make code always visible
options(scipen = 1)  # Turn off scientific notations for numbers
```

3. Loading the libraries and data

```r
library(datasets)
data("mtcars")
```

4. Exploratory Data Analysis and Summary of data

```r
# check the dataset structure
str (mtcars)
```

The dataset has 32 observations with 11 variables

```r
#check the header rows
head (mtcars)
```

Let's convert the important numeric variables into factors.

```r
#convert the numeric variables into factors
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
mtcars$am <- factor(mtcars$am,labels=c('Automatic','Manual'))


# check the modified structure
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am  : Factor w/ 2 levels "Automatic","Manual": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ carb: Factor w/ 6 levels "1","2","3","4",..: 4 4 1 1 2 1 4 2 2 4 ...
```

Initially, we will plot and observe the relationships between all the variables of the dataset (**see Figure 1 in Appendix**).

From the plot, we notice that variables like cyl, disp, hp, drat, wt, vs and am seem to have some strong correlation with mpg. Since we are interested in the effects of car transmission type on miles per gallon, we will plot & observe the comparison of both the varaibles (**see Figure 2 in Appendix**).

This plot clearly depicts an increase in the mpg when the transmission is Manual. This data is further analyzed and discussed in regression analysis section by fitting a linear model.

5. Regression Analyis

In this section, we will build linear regression models using different variables in order to find the best fit and compare it with the base model which we have using anova. After model selection, we also perform analysis of residuals.

5.1 Model building and selection

By observing pairs plot, we can easily conclude that there are multiple variables having high correlation with mpg. Hence, We will build an initial model with all the variables as predictors, and perfom stepwise model selection to select significant predictors for the final model which will be the best model. This will be taken care by the step method which runs lm multiple times to build multiple regression models and select the best variables from them using both forward selection and backward elimination methods by the AIC algorithm.

```r
init_model <- lm(mpg ~ ., data = mtcars)
best_model <- step(init_model, direction = "both")
```

The best model obtained from the above computations consists of the variables, cyl, wt and hp as confounders and am as the independent variable. Details of the model are depicted below.

```r
summary(best_model)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## amManual     1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

From the above model details, we observe that the adjusted R^2 value is 0.84 which is the maximum obtained considering all combinations of variables. Thus, we can conclude that more than 84% of the variability is explained by the above model.

In the following section, we compare the base model with only am as the predictor variable and the best model which we obtained earlier containing confounder variables also.

```r
base_model <- lm(mpg ~ am, data = mtcars)
anova(base_model, best_model)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at the above results, the p-value obtained is highly significant and we reject the null hypothesis that the confounder variables cyl, hp and wt don't contribute to the accuracy of the model.

5.2 Residuals & Diagnostics

In this section, we shall study the residual plots of our regression model and also compute some of the regression diagnostics for our model to find out some interesting leverage points (often called as outliers) in the data set. Please refer to Residuals & Diagnostics - Outliers plots (**see Figure 3 in Appendix**) for further details

From the above plots, we can make the following observations,

- The points in the Residuals vs. Fitted plot seem to be randomly scattered on the plot and verify the independence condition.
- The Normal Q-Q plot consists of the points which mostly fall on the line indicating that the residuals are normally distributed.
- The Scale-Location plot consists of points scattered in a constant band pattern, indicating constant variance.
- There are some distinct points of interest (outliers or leverage points) in the top right of the plots.

We now compute some regression diagnostics of our model to find out these interesting leverage points as shown in the following section. We compute top three points in each case of influence measures.

```r
leverage <- hatvalues(best_model)
tail(sort(leverage),3)
```

```
##       Toyota Corona Lincoln Continental       Maserati Bora
##           0.2777872           0.2936819           0.4713671
```

```r
influential <- dfbetas(best_model)
tail(sort(influential[,6]),3)
```

```
## Chrysler Imperial          Fiat 128     Toyota Corona
##         0.3507458         0.4292043         0.7305402
```

Looking at the above cars, we notice that our analysis was correct, as the same cars are mentioned in the residual plots.

5.3 Inference

We also perform a t-test assuming that the transmission data has a normal distribution and we clearly see that the manual and automatic transmissions are significatively different.

```r
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
```

6 Conclusion

Based on the observations from our best fit model, we can conclude the following,

- Cars with Manual transmission get more miles per gallon mpg compared to cars with Automatic transmission. (1.8 adjusted by hp, cyl, and wt).
- mpg will decrease by 2.5 (adjusted by hp, cyl, and am) for every 1000 lb increase in wt.
- mpg decreases negligibly with increase of hp.
- If number of cylinders, cyl increases from 4 to 6 and 8, mpg will decrease by a factor of 3 and 2.2 respectively (adjusted by hp, wt, and am).

7 Appendix

Figure 1: Plot explaining relationship between all the variables

```r
pairs(mtcars)
```

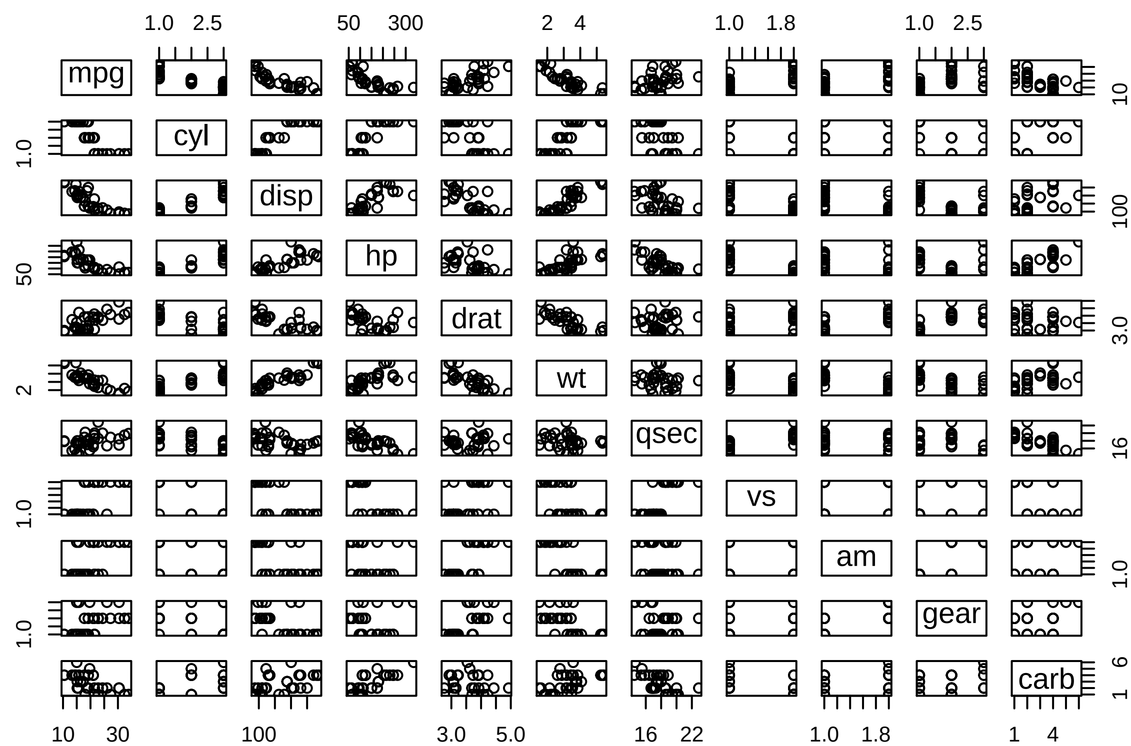


Figure 2: Boxplot explaining relationship between Miles per Gallon and Transmission Type

```r
boxplot(mpg ~ am, data = mtcars,
        col = (c("yellow","lightblue")),
        ylab = "Miles Per Gallon (mpg)",
        xlab = "Transmission Type (am)",
        main="Relation between mpg and Transmission Type")
```

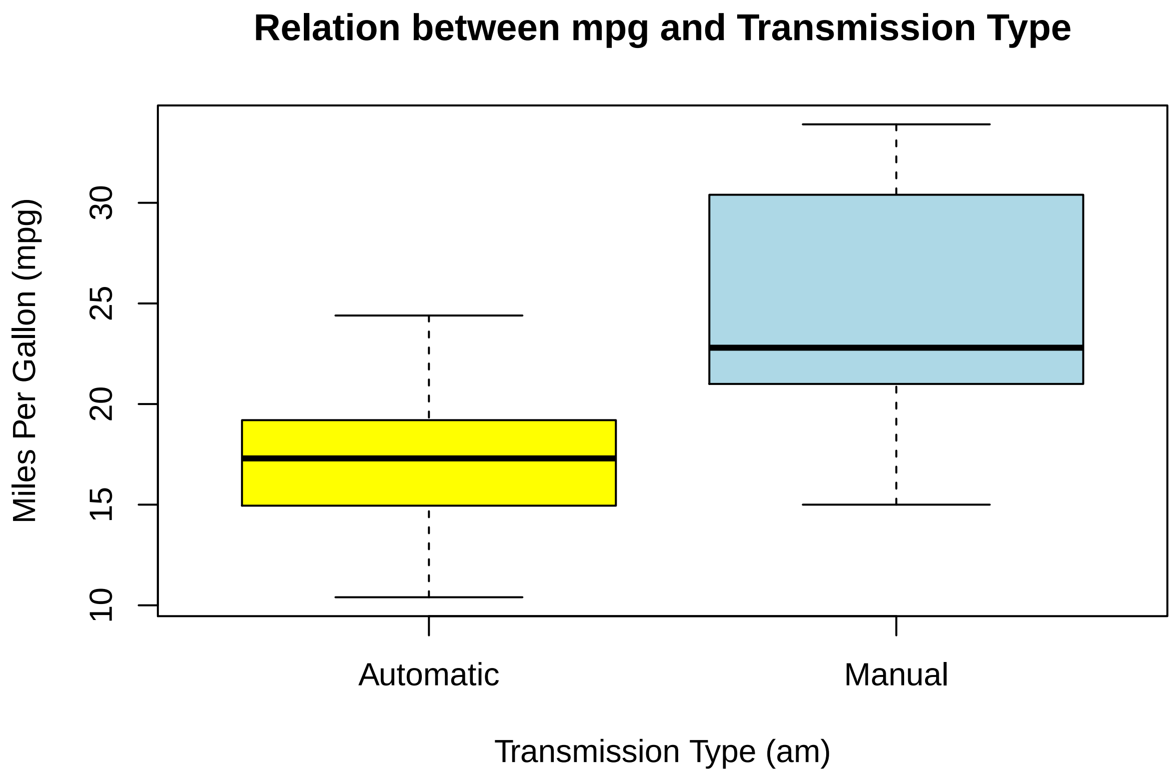
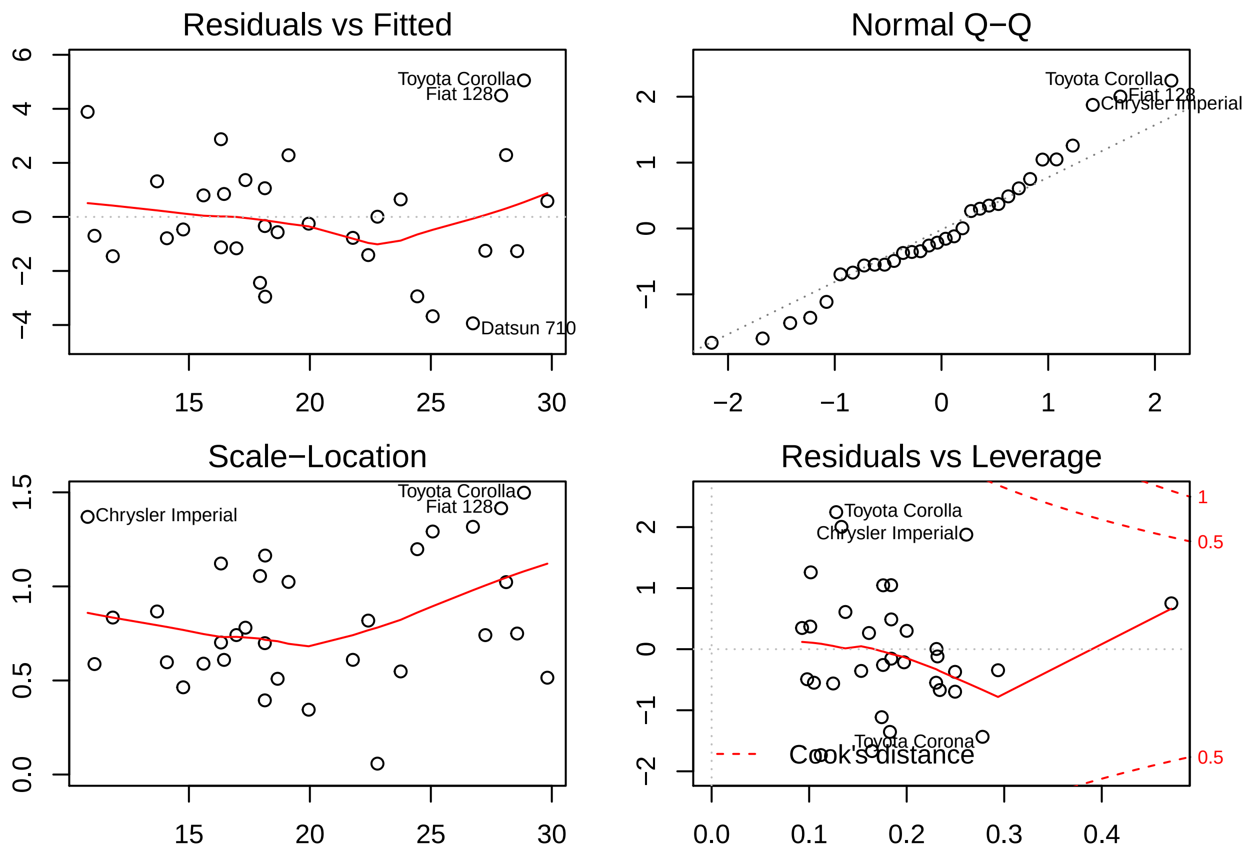


Figure 3 Residuals & Diagnostics - Outliers plots

```
par(mfrow = c(2, 2))
par(mar=c(2,2,2,2))
plot(best_model)
```



Residuals vs Fitted
Toyota Corolla
Fiat 128
Datsun 710
Normal Q−Q
Toyota Corolla
Fiat 128
Chrysler Imperial
Scale−Location
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals vs Leverage
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
1
0.5
0.5