Motor Trend Car Road Tests - Effects of transmission on MPG

Zhenkun Guo

March 27, 2016

1. Executive Summary

This detailed analysis has been 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 looks better for MPG but trend is not significant.
- MPG value for manual transmission is about 1.48 larger than for automatic transmission according to my best model

2. Data and Necessary Package

We load in the data set, perform the necessary data transformations by factoring the necessary variables and look at the data, in the following section.

```
library(ggplot2)
data("mtcars")
head(mtcars)
```

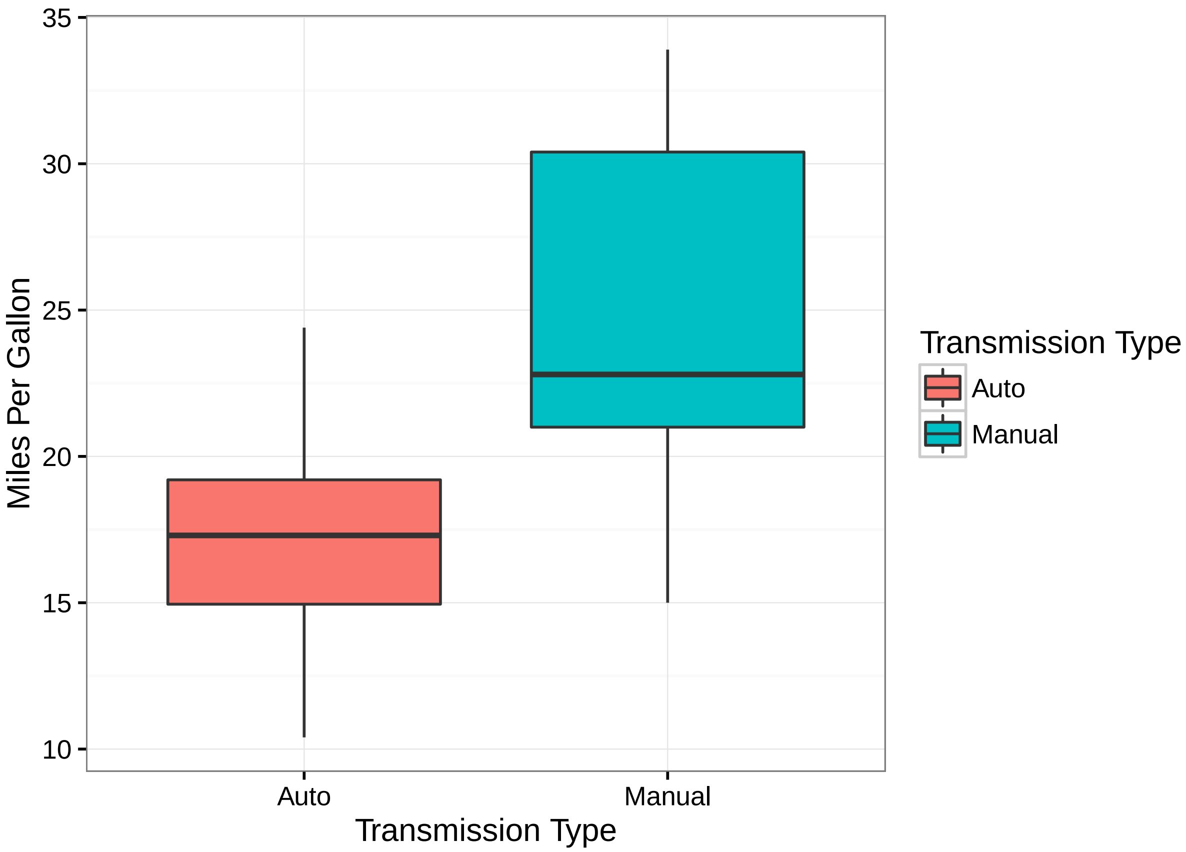
```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 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
```

3. Exploratory Data Analysis

3.1. Direct Comparasion of Transmission Types

Since we are interested in the effects of car transmission type on mpg, we plot boxplots of the variable mpg when am is Automatic or Manual. This plot depicts an increase in the mpg when the transmission is Manual.

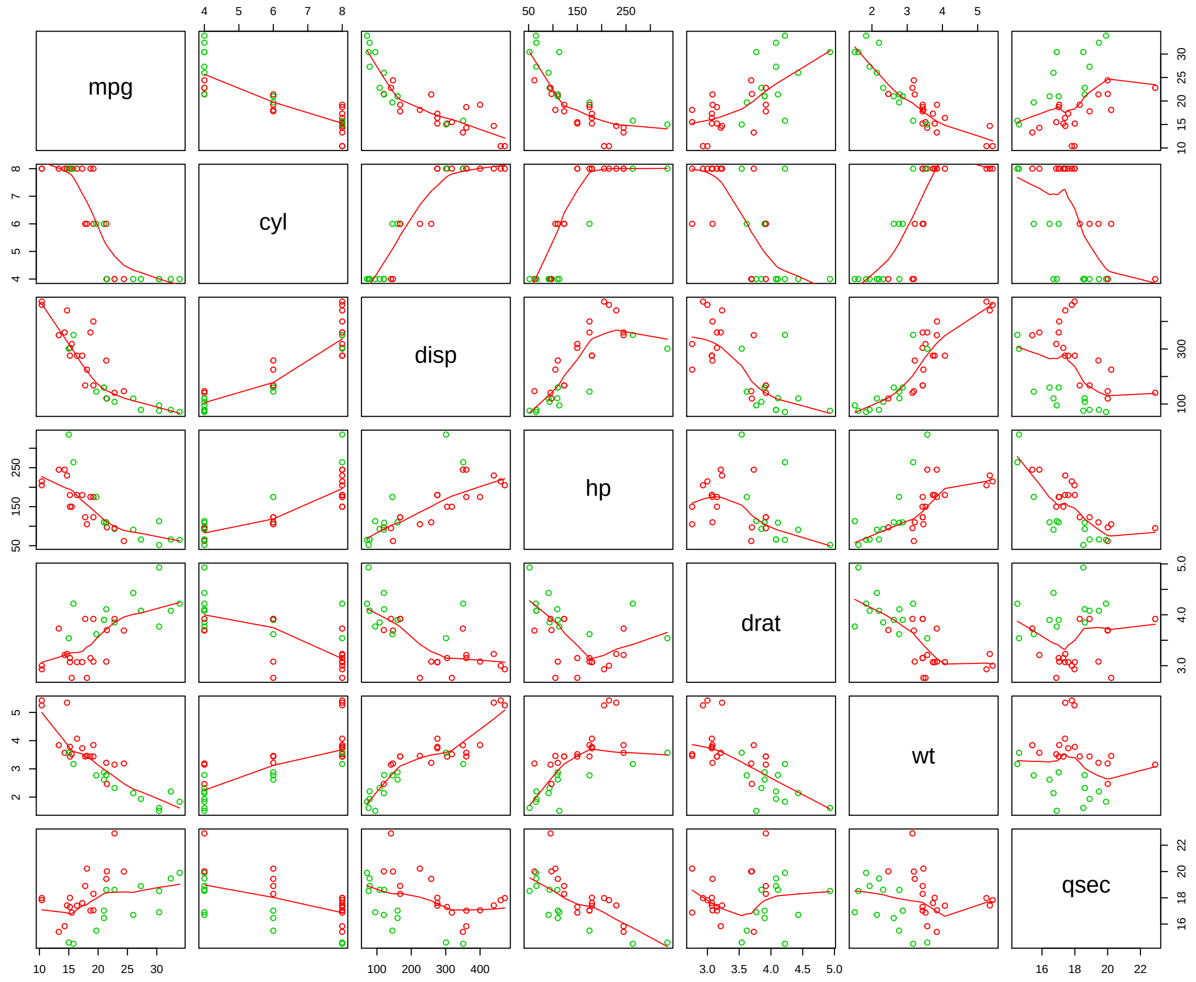
```
figure1<-ggplot(mtcars,aes(x=factor(am, labels = c("Auto","Manual")),y=mpg,fill=factor(am, labels = c("
figure1<-figure1+geom_boxplot()
figure1<-figure1+scale_fill_discrete(name = "Transmission Type")
figure1<-figure1 + theme_bw() + xlab("Transmission Type") + ylab("Miles Per Gallon")
figure1
```



3.2. Related Variables to MPG besides Transmission Type

In this section, we dive deeper into our data and explore various relationships between variables of interest. We plot the relationships between all the variables of the dataset. From the plot, we notice that variables like cyl, disp, hp, drat, wt, qsec seem to have some strong correlation with mpg. But we will use linear models to quantify that in the regression analysis section. In the figure, red dots indicate automatic transmission and green dots indicate manual transimission.

```
nt_mtcars<-subset(mtcars,select = c(1:7))
pairs(nt_mtcars, panel = panel.smooth, col =  mtcars$am +2)
```



4. Regression Analysis

4.1 Simple Model with Only Transmission Type

```
fit1<-lm(mpg ~ factor(am), data = mtcars)
summary(fit1)
```

```
##
## Call:
## lm(formula = mpg ~ factor(am), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## factor(am)1    7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

With this simple model, we can verify the idea than mpg for manual transmission is higher than automatic transmission by 7.24 if we ignore the influence of other variables. Adjusted R-squared is only 0.3385, which mean the model can't explain the variation of mpg well.

4.2. Four Regressors Model with AM, CYL, HP, WT

```
fit4<-lm(mpg ~ factor(am) + cyl + hp + wt, data = mtcars)
summary(fit4)
```

```
##
## Call:
## lm(formula = mpg ~ factor(am) + cyl + hp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4765 -1.8471 -0.5544  1.2758  5.6608
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.14654    3.10478  11.642 4.94e-12 ***
## factor(am)1  1.47805    1.44115   1.026   0.3142
## cyl         -0.74516    0.58279  -1.279   0.2119
## hp          -0.02495    0.01365  -1.828   0.0786 .
## wt          -2.60648    0.91984  -2.834   0.0086 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.509 on 27 degrees of freedom
## Multiple R-squared:  0.849,  Adjusted R-squared:  0.8267
## F-statistic: 37.96 on 4 and 27 DF,  p-value: 1.025e-10
```

With this model, we can see the mpg for manual transmission is higher than automatic by 1.48, which is much lower than the simpel model. This can be understood because the correlation between the transmission type and other regressors like cyl, hp or wt. This idea can be verified by the pairs plots seeing the distribution of colored dots. Adjusted R-squared is 0.8267.

4.3. Six Regressors Model with AM, CYL, HP, WT, DISP, QSEC

```
fit6<-lm(mpg ~ factor(am) + cyl + hp + wt + disp +qsec, data = mtcars)
summary(fit6)
```

```
##
## Call:
## lm(formula = mpg ~ factor(am) + cyl + hp + wt + disp + qsec,
##     data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6755 -1.6757 -0.4477  1.2615  4.6289
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.05170   13.30486   1.507  0.14432
## factor(am)1  2.94075    1.71810   1.712  0.09935 .
## cyl         -0.50207    0.78882  -0.636  0.53025
## hp          -0.01956    0.01489  -1.314  0.20088
## wt          -3.99773    1.21564  -3.289  0.00299 **
## disp         0.01396    0.01155   1.209  0.23802
## qsec         0.81018    0.57171   1.417  0.16879
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.458 on 25 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8337
## F-statistic: 26.91 on 6 and 25 DF,  p-value: 9.29e-10
```

With this model, we can see the mpg for manual transmission is higher than automatic by 2.94. Adjusted R-squared is 0.8337 slighter higher than the four regressors model.

5. Model Selection and Residual Analysis

5.1. Model Selection

The R-squared value is relatively low for the simplest model, so this model might not be enough to explain the variation of mpg. However, by adding two more regressors DISP and QSEC, the increasement of R-squared is not significant. So six regressors model might not be neccessary. To verifty this idea, we call the anova function to compare the models.

```
anova(fit1,fit4,fit6)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ factor(am) + cyl + hp + wt
## Model 3: mpg ~ factor(am) + cyl + hp + wt + disp + qsec
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     27 170.00  3    550.90 30.4046 1.674e-08 ***
## 3     25 150.99  2     19.01  1.5735    0.2272
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see adding cyl, hp and wt to the model is significant but adding two more regressors (disp, qsec) is not significant. So we decide here, the four regressors model with am, cyl, hp and wt is our best model.

5.2. Conclusion and Confidence Interval

```
summary(fit4)
```

```
##
## Call:
## lm(formula = mpg ~ factor(am) + cyl + hp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4765 -1.8471 -0.5544  1.2758  5.6608
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.14654    3.10478  11.642 4.94e-12 ***
## factor(am)1  1.47805    1.44115   1.026   0.3142
## cyl         -0.74516    0.58279  -1.279   0.2119
## hp          -0.02495    0.01365  -1.828   0.0786 .
## wt          -2.60648    0.91984  -2.834   0.0086 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.509 on 27 degrees of freedom
## Multiple R-squared:  0.849,  Adjusted R-squared:  0.8267
## F-statistic: 37.96 on 4 and 27 DF,  p-value: 1.025e-10
```

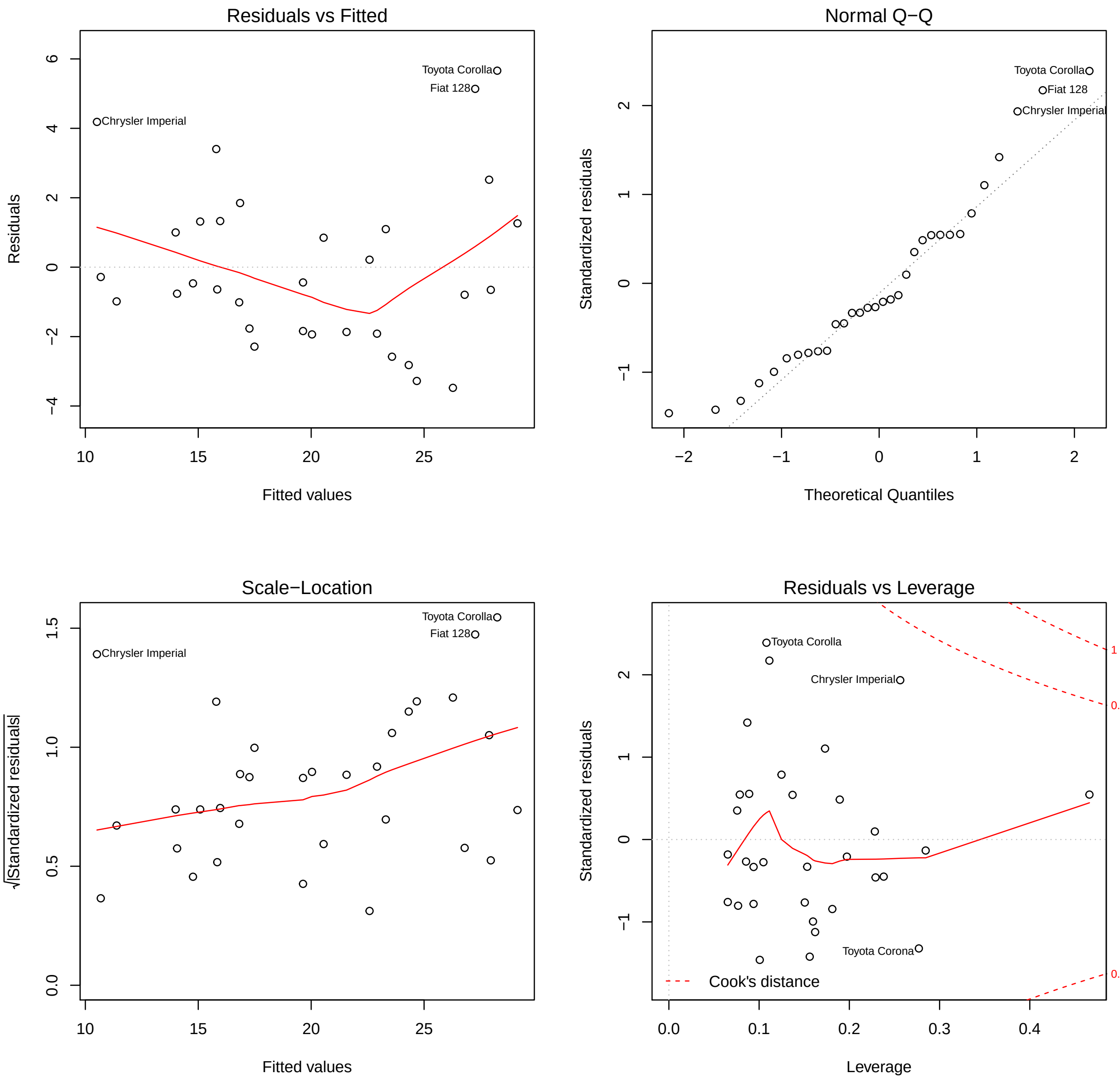
```
confint(fit4)
```

```
##                   2.5 %       97.5 %
## (Intercept) 29.77605177 42.517019733
## factor(am)1 -1.47894635  4.435041763
## cyl         -1.94093802  0.450623969
## hp          -0.05295064  0.003048517
## wt          -4.49383134 -0.719130075
```

P-value for am is 0.3142, which is not significant. We can also see here the 95% confidence interval for the influence of transmission type is (-1.48,4.44), which includes 0. So, here we conclude that manual transimission is not significantly better than automatic transmission respect to mpg.

5.3. Residual Analysis

```r
par(mfrow = c(2, 2))
plot(fit4)
```



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 (Toyota Carolla and Chrysler Imperial for instance) in the plots.

This model works well for the data