Motor Trend - A study of factors impacting MPG

Asif M Adnan

September 4, 2016

Executive Summary

In this article we will study "mtcars" data set and find miles per galon (MPG) dependency based on some other parameters. The given data set has `32` samples, and each of the samples have the follwing attributes, `mpg, cyl, disp, hp, drat, wt, qsec, vs, am, gear, carb`. The purpose of this article will be focused around,

- Is an automatic or manual transmission better for MPG?
- Quantify the MPG difference between automatic and manual transmissions?

In order to generate our report, we will go through the following steps one by one,

- Process and Explore target data set.
- Select a good model that can explain the relationship of MPG with the other factors, compared to some counterpart models (Selection and Comparison).
- Conclusion.

Data Process and Explore

Process

For this analysis, we need to load the following libraries,

```
library(ggplot2)
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 3.2.4
```

Before exploring and analyzing the data set, we need to load and process mtcars. Dimesion of the data set, `32, 11`. Attributes of the samples, `mpg, cyl, disp, hp, drat, wt, qsec, vs, am, gear, carb`.

Some samples

```
rbind(head(mtcars,3),tail(mtcars,3))
```

```
##                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
## Ferrari Dino  19.7   6  145 175 3.62 2.770 15.50  0  1    5    6
## Maserati Bora 15.0   8  301 335 3.54 3.570 14.60  0  1    5    8
## Volvo 142E    21.4   4  121 109 4.11 2.780 18.60  1  1    4    2
```

There are some attributes, which were used as factors (cyl, am, etc.). We need to convert the raw dataset into a format which will help us analyzing the data (code not shown for space).

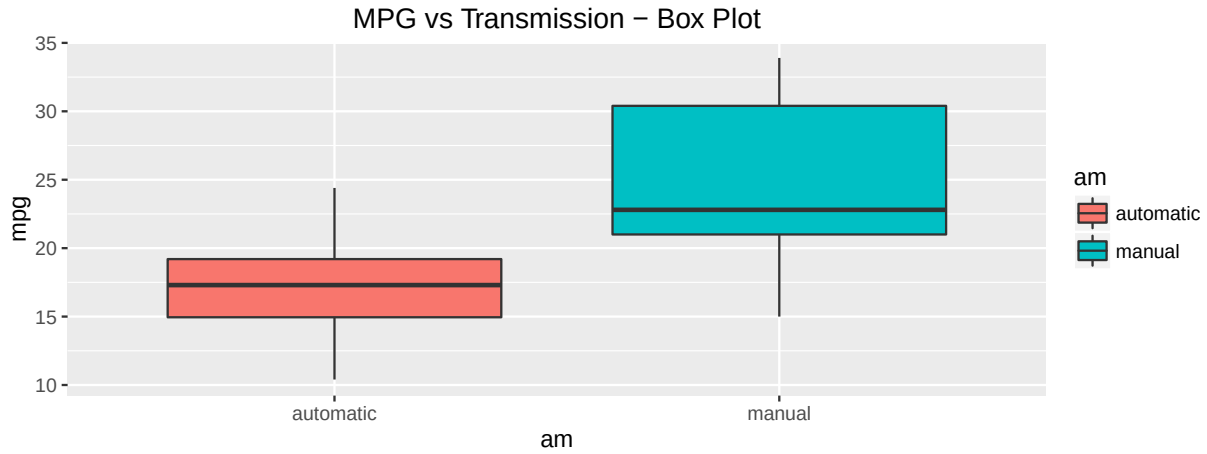
```
mtcarsModified <- mtcars
```

Explore

If we draw a box of the data based on the two variables we are interested (MPG vs Transmission), we'll get the following diagram,

```
mgpVStransmission <- ggplot(mtcarsModified, aes(am, mpg))+
                     geom_boxplot(aes(fill = am))+
                     labs(title = "MPG vs Transmission - Box Plot")

mgpVStransmission
```



The above graph is showing that, MPG for automatic and manual transmissions are significantly different. We can verify that from the following t-test.

```
mpgVStransmissionTest <- t.test(mtcarsModified$mpg~mtcarsModified$am)
mpgVStransmissionTest$p.value
```

```
## [1] 0.001373638
```

```
mpgVStransmissionTest$estimate
```

```
## mean in group automatic    mean in group manual
##                17.14737                24.39231
```

The p-value is siggesting that, automatic and manual transmission data are from two different populations. And the estimated means are quite similar to the box plot we drew.Even though these test suggest that automatic and manusal transmission cars have significantly different MPGs, in order to make sure that there are no biases in the data, we need to farther investigate.

Model Selection

We will use linear model to predict the outcome of MPG. At first, we will use only 'am' as the predictor. For the second model, we will use all the other variables. As for selecting the best model, we wil use R's step() function to decide the best model based on the model having all the variables.

Basic model having 'am' as the predictor - AM Model

We will use 'am' attribute as the predictor to decide MPG outcome.

```
fitAM <- lm(mpg~am, data = mtcarsModified)
#summary(fitAM)
```

According to this very basic model, an automatic transmission vehicle has on an average `17.1473684` MPG, where as, a manual tranmission vehicle has MPG increased by `7.2449393`. This model has Residual standard error as `4.9020288` on `30`, and the adjusted R-Squared value is `.3385`. R-squared value indicates that, this model can explain only approximately `33%` of the MPG variance.

All the attributes as the predictor - All Inclusion Model

In this model, we will use all the available attributes as the predictor.

```
fitAll <- lm(mpg~., data = mtcarsModified)
#summary(fitAll)
```

According to this all inclusive model, an automatic transmission vehicle has on an average `15.0926155` MPG, where as, a manual tranmission vehicle has MPG increased by `3.3473571`. This model has Residual standard error as `2.6162575` on `19`, and the adjusted R-Squared value is `.8116`. R-squared value indicates that, this model can explain approximately `81%` of the MPG variance, which is better than the 'am' model. Beside these, none of the coefficients are significant at 0.05 level.

Automatic selection of a better model - Auto Select Model

We will use R's step() function in order to select a comparatively good fit.

```
fitBetter <- step(fitAll, direction = "both", trace = FALSE)
#summary(fitBetter)
```

According to this automatically selected, better model, an automatic transmission vehicle has on an average `9.6177805` MPG, where as, a manual tranmission vehicle has MPG increased by `2.9358372`. This model has Residual standard error as `2.4588465` on `28`, and the adjusted R-Squared value is `.8336`. R-squared value indicates that, this model can explain approximately `83%` of the MPG variance, which is better than the all inclusive model.

Selecting an even better model - Best Model

We will use the attributes selected by the previous model, and multiply 'wt' and 'am' as another predictor.

```r
fitBest <- lm(mpg~wt + qsec + am + wt*am, data = mtcarsModified)
summary(fitBest)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am + wt * am, data = mtcarsModified)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5076 -1.3801 -0.5588  1.0630  4.3684
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.723      5.899   1.648 0.110893
## wt            -2.937      0.666  -4.409 0.000149 ***
## qsec           1.017      0.252   4.035 0.000403 ***
## ammanual      14.079      3.435   4.099 0.000341 ***
## wt:ammanual   -4.141      1.197  -3.460 0.001809 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.084 on 27 degrees of freedom
## Multiple R-squared:  0.8959, Adjusted R-squared:  0.8804
## F-statistic: 58.06 on 4 and 27 DF,  p-value: 7.168e-13
```

According to this automatically selected, better model, an automatic transmission vehicle has on an average `9.7230526` MPG, where as, a manual tranmission vehicle has MPG increased by `14.0794278`. This model has Residual standard error as `2.0841223` on `27`, and the adjusted R-Squared value is `.8804`. R-squared value indicates that, this model can explain approximately `88%` of the MPG variance, which is better than the previous one.

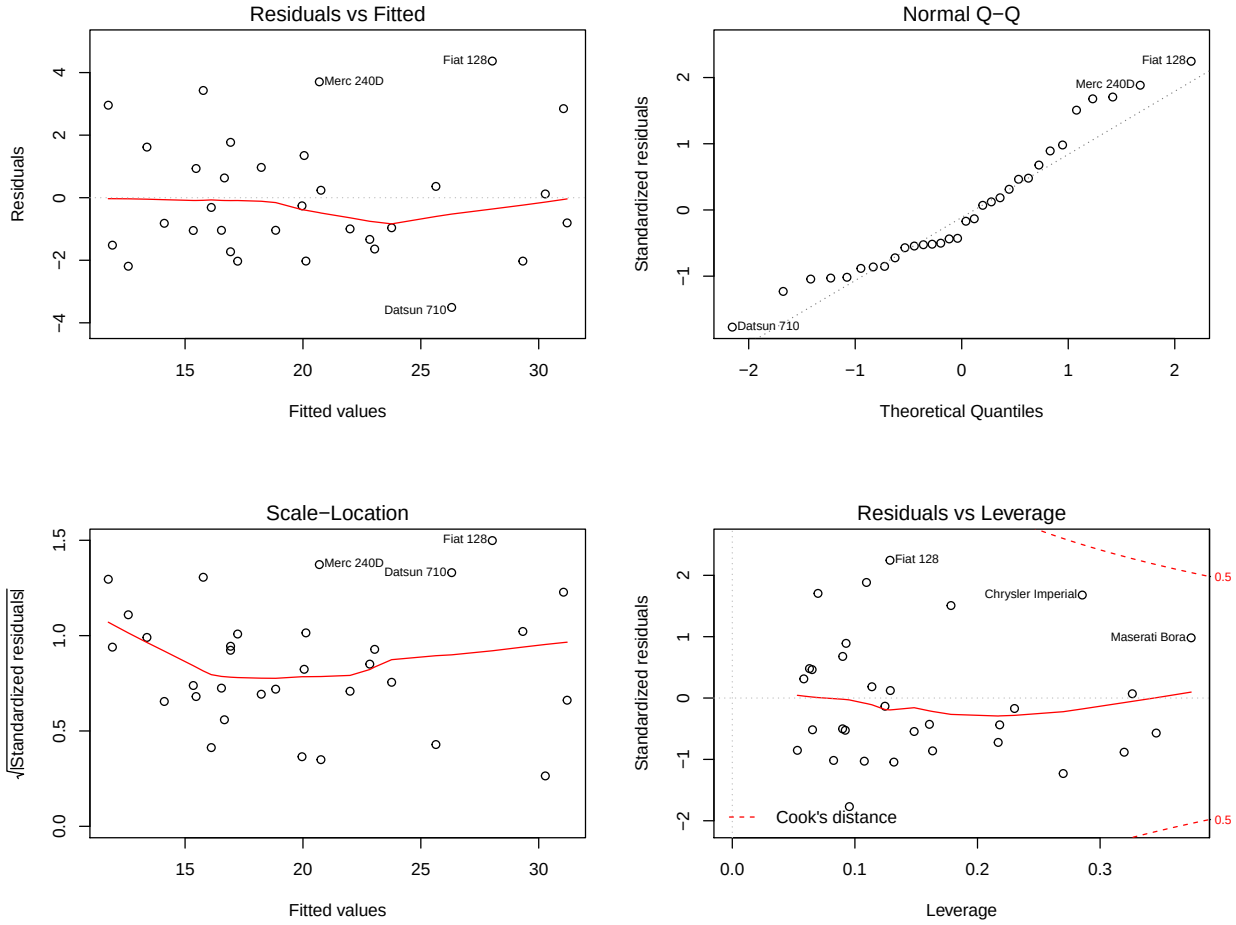
Comparison

Because of space, we did not print all the summaries, except the finally selected best one. The following table is displaying all the necessary attributes compared for the four model we discussed.

rows	AM_Model	All_Inclusion_Model	Auto_Select_Model	Best_Model
Avg Automatic MPG	17.15	15.09	9.62	9.72
Increase in Manual Avg MPG	7.24	3.35	2.94	14.08
Residual Standard Error	4.9	2.62	2.46	2.08
Degrees of Freedom	30	19	28	27
Adjusted R-squared	0.33	0.81	0.83	0.88
Precentage Explained	33	81	83	88

We will plot the best selected model in the following diagram.

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



According to the plot, there is consistent pattern observed (based on Fitted vs Residual plot). The residuals are normally distributed. The scale-location is showing that, points are randomly distributed. Also, there was no observable outlier shown in the Residue-Leverage plot.

Conclusion

We started our analysis with two questions in mind. The firs one is, "Is an automatic or manual transmission better for MPG?". Based on the models that we have discussed, it is quite obvious and easy to answer that, a manual transmission is better than an automatic one for MPG. Obviously, there are some certain chances that for a sample not present in this data set, an automatic transmission vahicle might have better MPG values. But with the given set of data, our best model can explain `88%` of the variances.

The secnond question in our list is, "Quantify the MPG difference between automatic and manual transmissions". This is not as straightforwad as the firs one is. If we look the comparison table given in the previous section, for different model, the difference in MPG values for manual vs automatic vehicles varies from model to model. The very basic 'am' model has a difference of `7.24` with low confidence, where as the best model has `14.08` differnce with high confidence. We also have another model (automatically selected one) which shows a differnce of `2.94` but with high confidence. If we analyze the Residual vs Fitted plots of the models (except for the basic 'am' one), we can see some randomness in the given data. These models also has large confidence values. But with only 32 samples, it is not very wise to generalize the notion that, a manual vehicle will always have migher MPG values comparing to the automatic counterparts.