The Impact that the Type of Transmission has on Fuel Efficiency

Trenton Potgieter

Saturday, April 18, 2015

Executive Summary

One of the key criteria in any vehicle purchasing decision is fuel economy, or how many Miles per Gallon (MPG). Added to this, there are various additional factors that need to be taken into consideration when assessing the effects on fuel-economy. Among these is the type of car transmission the vehicle is equipped with. Which of the two types of transmissions (**manual** or **automatic**) impacts fuel economy more?

This report will answer this question by analyzing what impact each type of transmission has on fuel economy and furthermore, the report will quantify the impact of each on the respective Miles per Gallon (MPG).

Data Processing

The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).

Initial Analysis

```r
#Summaries
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 : num  6 6 4 6 8 6 8 4 4 6 ...
 $ 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  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

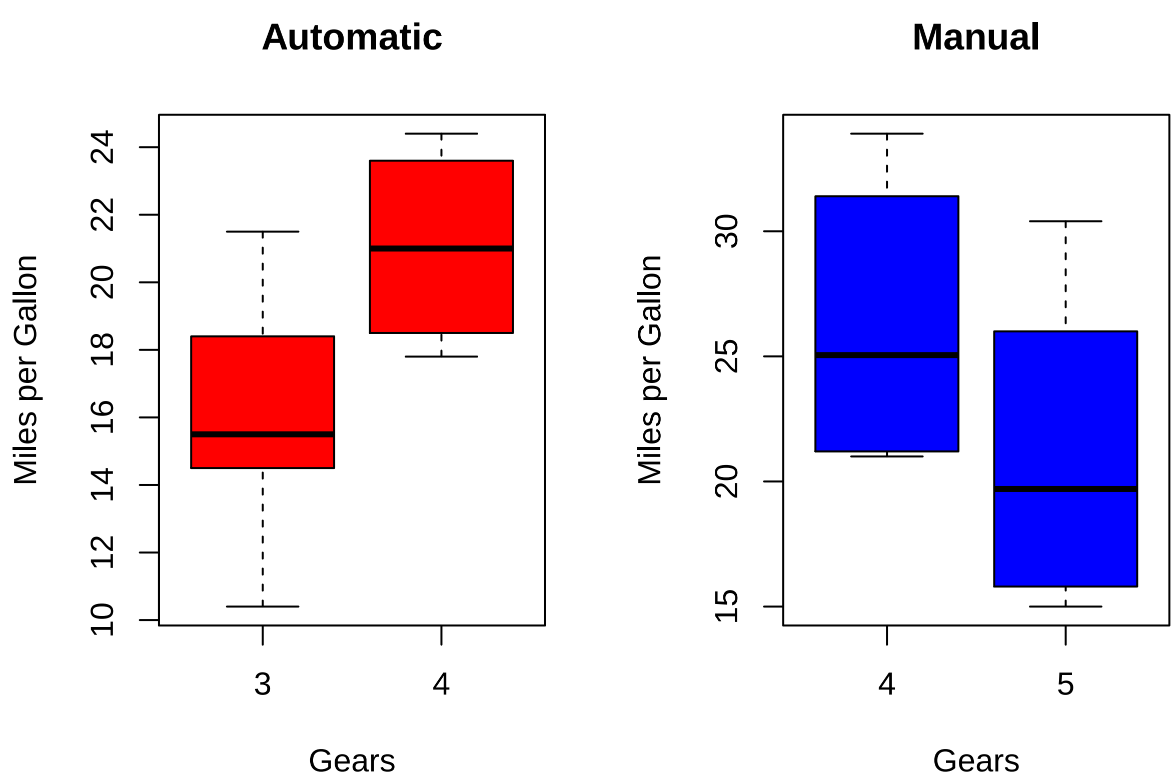
```r
summary(mtcars)
```

```
      mpg             cyl             disp             hp             drat             wt
 Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0   Min.   :2.760   Min.   :1.513
 1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080   1st Qu.:2.581
 Median :19.20   Median :6.000   Median :196.3   Median :123.0   Median :3.695   Median :3.325
 Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7   Mean   :3.597   Mean   :3.217
```

```
 3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920   3rd Qu.:3.610  
 Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0   Max.   :4.930   Max.   :5.424  
      qsec             vs               am              gear            carb      
 Min.   :14.50   Min.   :0.0000   Min.   :0.0000   Min.   :3.000   Min.   :1.000  
 1st Qu.:16.89   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
 Median :17.71   Median :0.0000   Median :0.0000   Median :4.000   Median :2.000  
 Mean   :17.85   Mean   :0.4375   Mean   :0.4062   Mean   :3.688   Mean   :2.812  
 3rd Qu.:18.90   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
 Max.   :22.90   Max.   :1.0000   Max.   :1.0000   Max.   :5.000   Max.   :8.000  
```

As can be seen from the summaries above, the data is made up of **32** separate motor vehicles, with **r ncol(mtcars)** different features. Since the focus of this report is on the MPG and Transmission type, there are **19** cars with **Automatic** transmissions and **13** with **Manual** transmissions.

```r
#Box Plot
mean_mpg <- as.list(round(tapply(mtcars$mpg, mtcars$am, mean), 2))
par(mfrow = c(1, 2))
boxplot(mtcars$mpg ~ mtcars$gear, subset = (mtcars$am == 0),
        xlab = "Gears", ylab = "Miles per Gallon", col = "red",
        main = "Automatic")
boxplot(mtcars$mpg ~ mtcars$gear, subset = (mtcars$am == 1),
        xlab = "Gears", ylab = "Miles per Gallon", col = "blue",
        main = "Manual")
```



The above plot shows the distribution of the transmission types for the various vehicles and their respective MPG. As can be seen, the vehicles with a Manual transmission have an average of **24.39** MPG, while the vehicles with an Automatic transmission have an average of **17.15** MPG. It also seems that the vehicles with **4** gears (whether Automatic or Manual) are more economical. Since there are other variables that can influence this conclusion (See Appendix A: Pair-wise Plot for other highly correlated variables), the rest of this report will quantify these more.

Data Transformations

There are no missing values or outliers within the data set, however there are a number of variables that have a number of categories. So to prepare the data for a better analysis, we factorize the variables into their individual categories and provide better names for the type of transmission.

```
#Factorize the variables with catagories
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- as.factor(mtcars$am)
levels(mtcars$am) <- c("Auto", "Man") #Rename Transmission Type
mtcars$gear <- as.factor(mtcars$gear)
mtcars$carb <- as.factor(mtcars$carb)
```

Regression Analysis

Now that the data is neatened and factorized, we can run an initial regression analysis to get an idea of which of the variables have an effect on fuel economy. The first model to fit in the initial assumption that just the type of transmission affects fuel economy.

```
#Fit the model
fit <- lm(mpg~am, data = mtcars)
summary(fit)
```

```
Call:
lm(formula = mpg ~ 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 ***
amMan          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
```

As can be seen from the output above, the original assumption shows that the **Manual** transmission type significantly influences the **MPG**. But as mentioned at the outset, what about the other variables?

Model Selection

To start the process of finding the best model we fit the linear regression model using all the variables as predictors.

```
#Fit a regression model on ALL the data
full_fit <- lm(mpg~., data = mtcars)

#Show the highest corrleated coeficients
summary(full_fit)
```

```
Call:
lm(formula = mpg ~ ., data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5087 -1.3584 -0.0948  0.7745  4.6251

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.87913   20.06582   1.190   0.2525
cyl6        -2.64870    3.04089  -0.871   0.3975
cyl8        -0.33616    7.15954  -0.047   0.9632
disp         0.03555    0.03190   1.114   0.2827
hp          -0.07051    0.03943  -1.788   0.0939 .
drat         1.18283    2.48348   0.476   0.6407
wt          -4.52978    2.53875  -1.784   0.0946 .
qsec         0.36784    0.93540   0.393   0.6997
vs1          1.93085    2.87126   0.672   0.5115
amMan        1.21212    3.21355   0.377   0.7113
gear4        1.11435    3.79952   0.293   0.7733
gear5        2.52840    3.73636   0.677   0.5089
carb2       -0.97935    2.31797  -0.423   0.6787
carb3        2.99964    4.29355   0.699   0.4955
carb4        1.09142    4.44962   0.245   0.8096
carb6        4.47757    6.38406   0.701   0.4938
carb8        7.25041    8.36057   0.867   0.3995
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.833 on 15 degrees of freedom
Multiple R-squared:  0.8931,    Adjusted R-squared:  0.779
F-statistic:  7.83 on 16 and 15 DF,  p-value: 0.000124
```

The above output shows the list of variables that have the most influence on fuel economy. When compared with the Initial Analysis and initial model fit, we can now see the list of coefficients that have some impact on fuel economy, not just the type of transmission. Since the objective of this report is to determine the most significant factors that influence fuel economy. It is clear from the output that there are still too many variables, namely **16**. Therefore, it is prudent to find a better fitting model. To this end, we make use of the `step()`[1] function in `R` to make use of a search algorithm[2] to find the best model. Automatic methods

[1] http://www.stat.columbia.edu/~martin/W2024/R10.pdf
[2] http://web.mit.edu/r_v3.0.1/lib/R/library/stats/html/step.html

are useful when the number of explanatory variables is large and it is not feasible to fit all possible models. In this case we will use both a Forward selection as well as Backward elimination for model selection.

```
#Find the best model using the step() funciton
best_fit <- step(full_fit, direction = "both")
```

```
Start:  AIC=76.4
mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb

       Df Sum of Sq    RSS    AIC
- carb  5   13.5989 134.00 69.828
- gear  2    3.9729 124.38 73.442
- am    1    1.1420 121.55 74.705
- qsec  1    1.2413 121.64 74.732
- drat  1    1.8208 122.22 74.884
- cyl   2   10.9314 131.33 75.184
- vs    1    3.6299 124.03 75.354
<none>              120.40 76.403
- disp  1    9.9672 130.37 76.948
- wt    1   25.5541 145.96 80.562
- hp    1   25.6715 146.07 80.588

Step:  AIC=69.83
mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear

       Df Sum of Sq    RSS    AIC
- gear  2    5.0215 139.02 67.005
- disp  1    0.9934 135.00 68.064
- drat  1    1.1854 135.19 68.110
- vs    1    3.6763 137.68 68.694
- cyl   2   12.5642 146.57 68.696
- qsec  1    5.2634 139.26 69.061
<none>              134.00 69.828
- am    1   11.9255 145.93 70.556
- wt    1   19.7963 153.80 72.237
- hp    1   22.7935 156.79 72.855
+ carb  5   13.5989 120.40 76.403

Step:  AIC=67
mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am

       Df Sum of Sq    RSS    AIC
- drat  1    0.9672 139.99 65.227
- cyl   2   10.4247 149.45 65.319
- disp  1    1.5483 140.57 65.359
- vs    1    2.1829 141.21 65.503
- qsec  1    3.6324 142.66 65.830
<none>              139.02 67.005
- am    1   16.5665 155.59 68.608
- hp    1   18.1768 157.20 68.937
+ gear  2    5.0215 134.00 69.828
- wt    1   31.1896 170.21 71.482
+ carb  5   14.6475 124.38 73.442
```

```
Step:  AIC=65.23
mpg ~ cyl + disp + hp + wt + qsec + vs + am

       Df Sum of Sq    RSS    AIC
- disp  1    1.2474 141.24 63.511
- vs    1    2.3403 142.33 63.757
- cyl   2   12.3267 152.32 63.927
- qsec  1    3.1000 143.09 63.928
<none>              139.99 65.227
+ drat  1    0.9672 139.02 67.005
- hp    1   17.7382 157.73 67.044
- am    1   19.4660 159.46 67.393
+ gear  2    4.8033 135.19 68.110
- wt    1   30.7151 170.71 69.574
+ carb  5   13.0509 126.94 72.095

Step:  AIC=63.51
mpg ~ cyl + hp + wt + qsec + vs + am

       Df Sum of Sq    RSS    AIC
- qsec  1     2.442 143.68 62.059
- vs    1     2.744 143.98 62.126
- cyl   2    18.580 159.82 63.466
<none>              141.24 63.511
+ disp  1     1.247 139.99 65.227
+ drat  1     0.666 140.57 65.359
- hp    1    18.184 159.42 65.386
- am    1    18.885 160.12 65.527
+ gear  2     4.684 136.55 66.431
- wt    1    39.645 180.88 69.428
+ carb  5     2.331 138.91 72.978

Step:  AIC=62.06
mpg ~ cyl + hp + wt + vs + am

       Df Sum of Sq    RSS    AIC
- vs    1     7.346 151.03 61.655
<none>              143.68 62.059
- cyl   2    25.284 168.96 63.246
+ qsec  1     2.442 141.24 63.511
- am    1    16.443 160.12 63.527
+ disp  1     0.589 143.09 63.928
+ drat  1     0.330 143.35 63.986
+ gear  2     3.437 140.24 65.284
- hp    1    36.344 180.02 67.275
- wt    1    41.088 184.77 68.108
+ carb  5     3.480 140.20 71.275

Step:  AIC=61.65
mpg ~ cyl + hp + wt + am

       Df Sum of Sq    RSS    AIC
<none>              151.03 61.655
- am    1     9.752 160.78 61.657
```

```
+ vs    1     7.346 143.68 62.059
+ qsec  1     7.044 143.98 62.126
- cyl   2    29.265 180.29 63.323
+ disp  1     0.617 150.41 63.524
+ drat  1     0.220 150.81 63.608
+ gear  2     1.361 149.66 65.365
- hp    1    31.943 182.97 65.794
- wt    1    46.173 197.20 68.191
+ carb  5     5.633 145.39 70.438
```

```r
#Show the coefficents of the best fit
summary(best_fit)
```

```
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 ** 
amMan        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
```

The output shows that the coefficients that have the most impact on MPG. So by using automatic methods provides the best linear function.

Results

The regression analysis shows the following significant information:

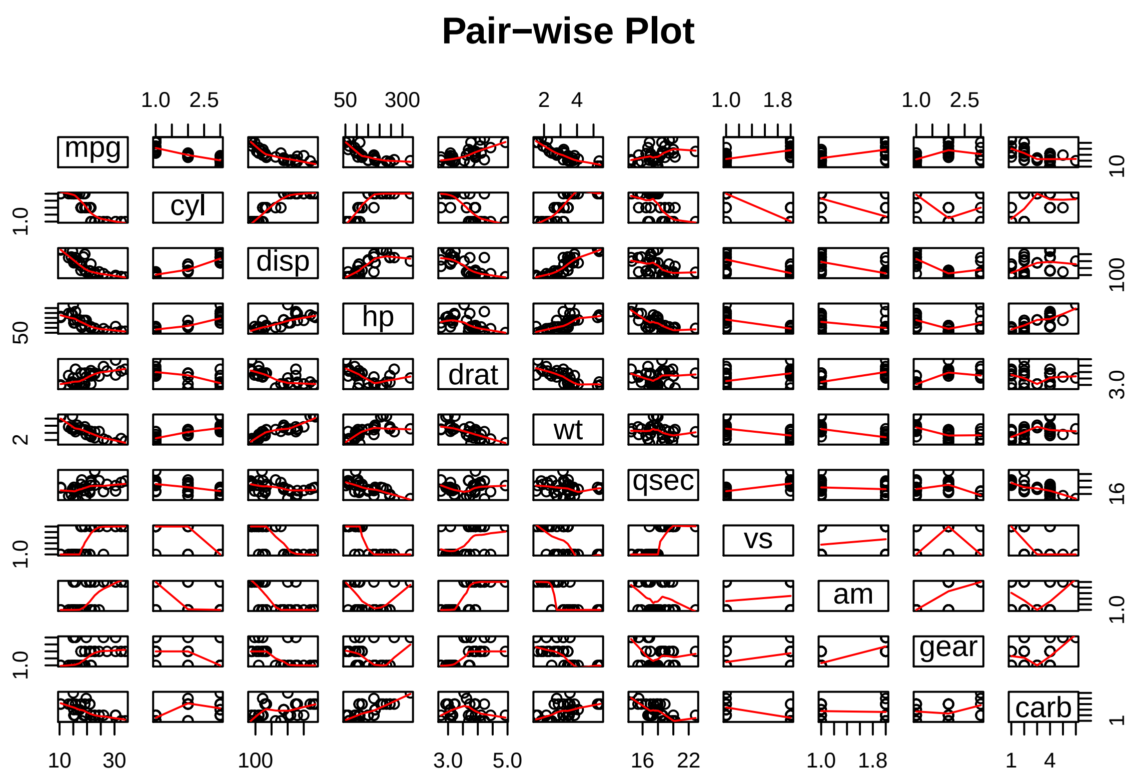
- The model used for the initial assumption produces an adjusted R^2 of **0.34**.
- The model used to fit all the variables as predictors produces an adjusted R^2 of **0.78**.
- The best model determined using automatic methods produces as adjusted R^2 of **0.84**. This shows that it certainly doesn't have the highest influence on fuel economy.

Conclusion

So not only does the **Manual** transmission impact the fuel economy of a vehicle, but as this report shows, so does having **6** and **8** cylinders, the **horsepower** and the **weight**.

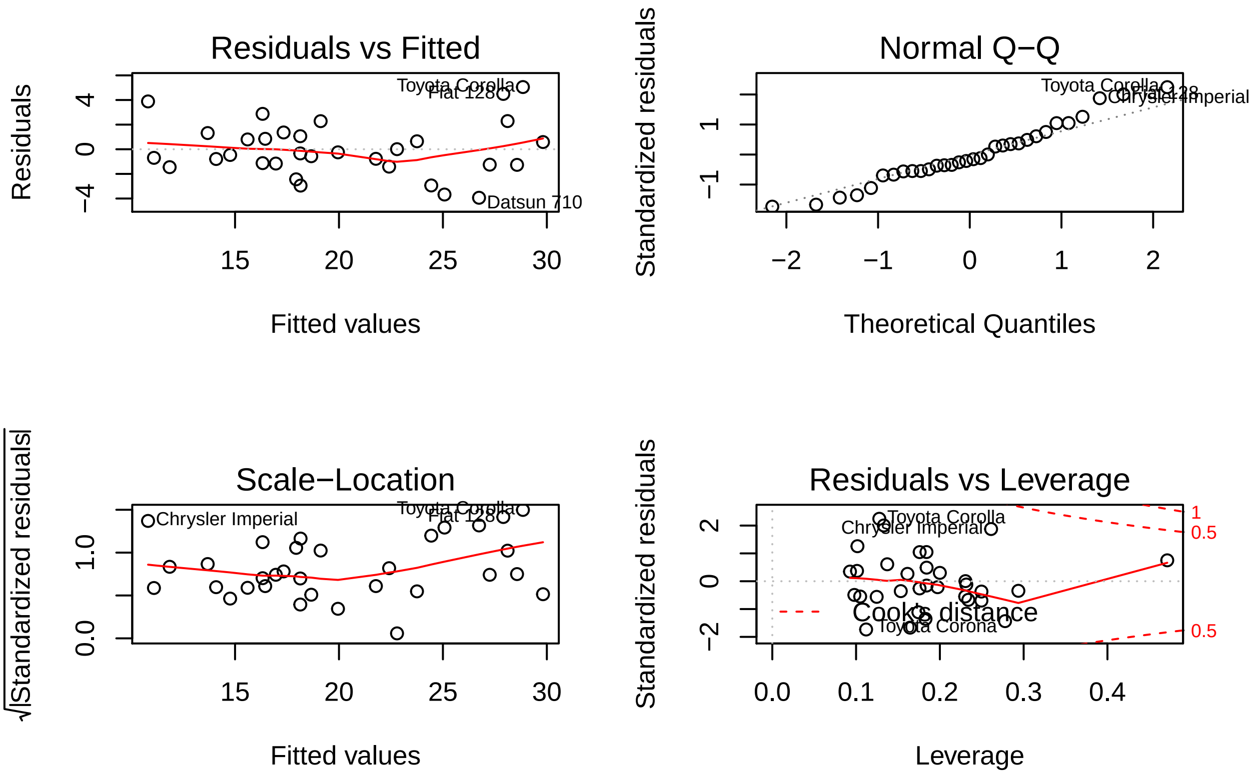
Appendix A: Pair-wise Plot

```r
#Pair-wise plot
pairs(mtcars, panel=panel.smooth, main="Pair-wise Plot")
```



Appendix B: Residual Plots

```r
#Plot the best fit
par(mfrow = c(2,2))
plot(best_fit)
```


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Datsun 710
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Chrysler Imperial
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Leverage
Cook's distance
Toyota Corona