Fuel Efficiency & Transmisison - A Deeper Analysis

Nagesh Madhwal

November 22, 2015

SUMMARY

Analysis of data from 32 car models indicates that it is not possible to conclusively state that manual transmission cars have higher fuel efficiency compared to automatic cars or vice versa. At the same time analysis also shows that the best model to explain the variability in the mileage of cars will include transmission type as a predictor. Exploratory data analysis shows us that manual transmission cars have a higher median fuel efficiency compared to automatic transmission cars but when we model the variability we see that the 95% confidence interval for impact of transmission type on mileage ranges from a negative 0.59 to a positive 5.996 miles per gallon & the pvalue is 0.1032. Presence of transmission in the best fit model shows that it does play a part in determining the variability but as a supporting variable to more significant predictors like weight & horse power.

Article

Big powerful cars, high weight, high horsepower, high displacement are gas guzzlers, we all know this from experience & on those factors our choices on adrenalin highs verses efficiency are quite easily made. When it comes to transmission though, the choice is not so clear. With improvements in technology automatic transmission cars have started to catch up with Manual ones on mileage. The convenience of Automatic can outweigh the loss in economy if the difference is not too high. So as an educated buyer we need to know whether the transmission type actually has a significant impact on the fuel economy and if it does how big is the hole we are burning in our pockets to purchase the convenience of automatic transmission? Motor trend magazine studied data for 32 car models over multiple parameters to try & answer this question.

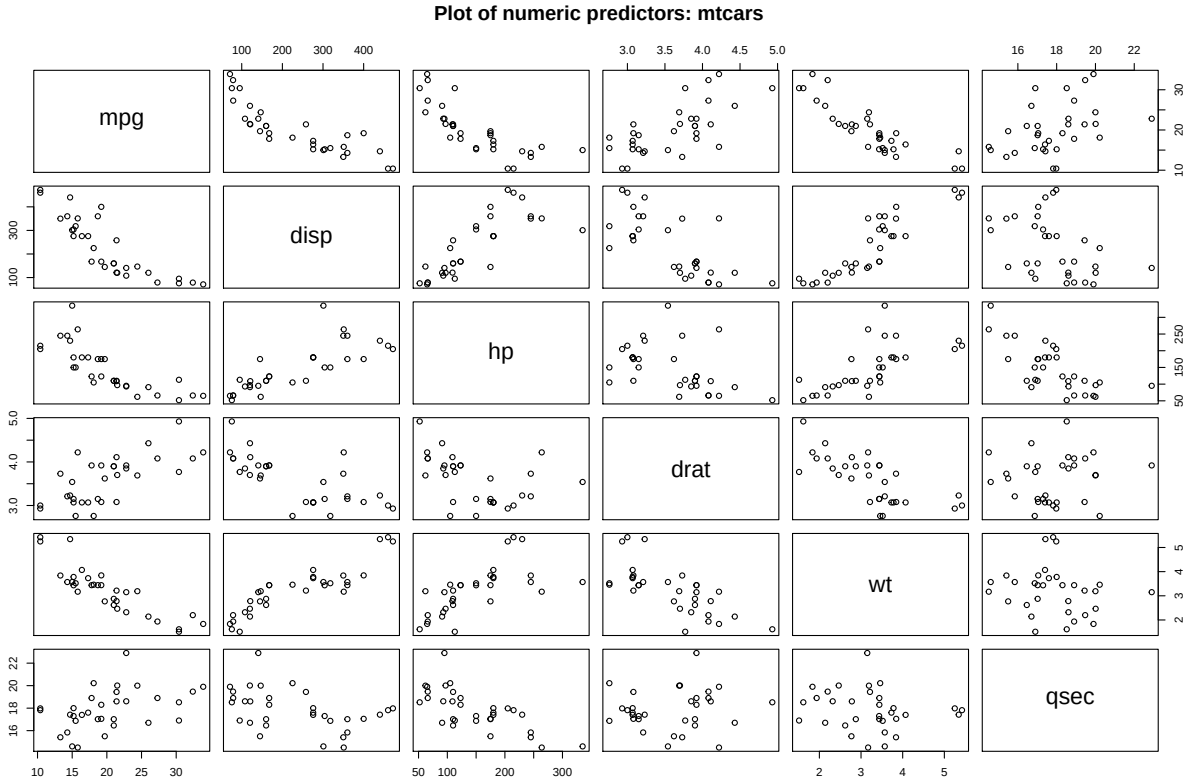
Let us first explain the methodology for the folks who are inclined more towards an understanding of how we did this analysis. We built our model using multiple regressions - a technique which allows us to build models between values of interest & their predictors. The assumption is a simple linear relationship and we use this approach as it allows us to better understand the impact & significance of the various predictors. As a first step we start by looking at the data & assess any initial trends and relationships that give us insights into building a model. We then start by building a model with all the variables we have in our dataset & we use a "backward selection" approach to drop indicators (say displacement, or number of gears) to refine the model. Our attempt is to reach the minimum number of predictors that explain the largest portion of the variability in the mileage of the car within the confines of a linear relationship. We also check the relevance of the model by seeing whether any of the unexplained aspects (which in technical terms are called "Residuals") have an identified trend which would mean that there still are aspects of the variability that the model is not explaining. Finally we also checked the confidence level in our model. The details of the approach are appended at the end of the article for people who want to immerse themselves into more details of the data.

A quick look at our data tells us some obvious knowns – weight, displacement, horsepower have a very direct negative relationship while rear axle ratio, and quarter mile time have a positive relationship with fuel efficiency. These variables are also very heavily correlated to each other and again we understand that intuitively, bigger heavier cars have more cylinders, more horsepower, and more displacement! This poses a challenge for us in studying the correct impact of the specific variables, its something we have to keep in mind as we build our model. Generally there is a decreasing trend in mileage as number of cylinders goes up. We see that the median mileage for manual cars is higher than the median mileage for automatics.

Model selection is done to arrive at the least number of predictors which would be able to explain the largest portion of variability in fuel efficiency. In an iterative process starting with the complete set of variables & working backwards we selected a model consisting of number of cylinders, weight, horsepower, engine

placement and transmission type. This came out as a significant model & of all the models we considered it captures the highest level of variablity in car fuel efficiency. Interestingly the only significant predictors in this model are weight and horsepower with a clearly negative relationship with miles per gallon. No of cylinders, engine placement (V vs Straight) and most important to our current interest transmission type are NOT significant predictors though they contribute to explaining the variability. Put more simply if we do not add transmission type to the linear regression model it will be less accurate in explaining the variability in mileage, but from the model we cannot definitely conclude that the manual transmission cars have a better mileage than automatic transmission cars.
The key parameters we considered in model selection were values for "Adjusted R squared"" & "Predicted R Squared". "Adjusted R squared"" gives us a measure of the amount of variability that is explained by the included predictors while "predicted R squared"" tells us how well the model will predict new observations or whether the model is too complicated. The model we selected has an "adjusted R squared" value of 0.84 & a "predicted R squared"" value of 0.805. None of the other models created are able to exceed these values. A point to note here is that just because a predictor is not significant it does not mean that it should be removed from the model. In this particular case non-significant predictors are contributing to explain the variability in the model. Checking the residuals plot we validate that there is no clear trend there, the values are randomly scattered. The output of our model tells us that all other things being equal on average manual transmission car would give an additional 2.7 miles per gallon. The 95% confidence interval is from " -0.588 (negative) to 5.996" miles per gallon. The negative value at the lower end along with the 10.32% pvalue indicate to us that we cannot conclude that manual transmission always results in a higher mileage compared to automatic.

ANNEXURE




```
##              mpg       disp         hp        drat         wt        qsec
## mpg   1.0000000 -0.8475514 -0.7761684  0.68117191 -0.8676594  0.41868403
```

```
## disp -0.8475514  1.0000000  0.7909486 -0.71021393  0.8879799 -0.43369788
## hp   -0.7761684  0.7909486  1.0000000 -0.44875912  0.6587479 -0.70822339
## drat  0.6811719 -0.7102139 -0.4487591  1.00000000 -0.7124406  0.09120476
## wt   -0.8676594  0.8879799  0.6587479 -0.71244065  1.0000000 -0.17471588
## qsec  0.4186840 -0.4336979 -0.7082234  0.09120476 -0.1747159  1.00000000

## Start from a model based on all predictors & work backwards

##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + qsec + vs + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2252 -1.4245 -0.1423  1.0696  4.2077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.86519   11.87783   2.009   0.0559 .
## cyl6        -1.63088    1.79566  -0.908   0.3728
## cyl8         0.82860    3.28782   0.252   0.8032
## hp          -0.02907    0.01654  -1.758   0.0915 .
## wt          -2.72220    1.04881  -2.596   0.0159 *
## qsec         0.41314    0.64135   0.644   0.5256
## vs1          1.37831    2.01846   0.683   0.5012
## am1          3.11177    1.73708   1.791   0.0859 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.426 on 24 degrees of freedom
## Multiple R-squared:  0.8746, Adjusted R-squared:  0.838
## F-statistic: 23.91 on 7 and 24 DF,  p-value: 2.345e-09

##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + vs + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3405 -1.2158  0.0046  0.9389  4.6354
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 31.18461    3.42002   9.118    2e-09 ***
## cyl6        -2.09011    1.62868  -1.283   0.2112
## cyl8         0.29098    3.14270   0.093   0.9270
## hp          -0.03475    0.01382  -2.515   0.0187 *
## wt          -2.37337    0.88763  -2.674   0.0130 *
## vs1          1.99000    1.76018   1.131   0.2690
## am1          2.70384    1.59850   1.691   0.1032
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

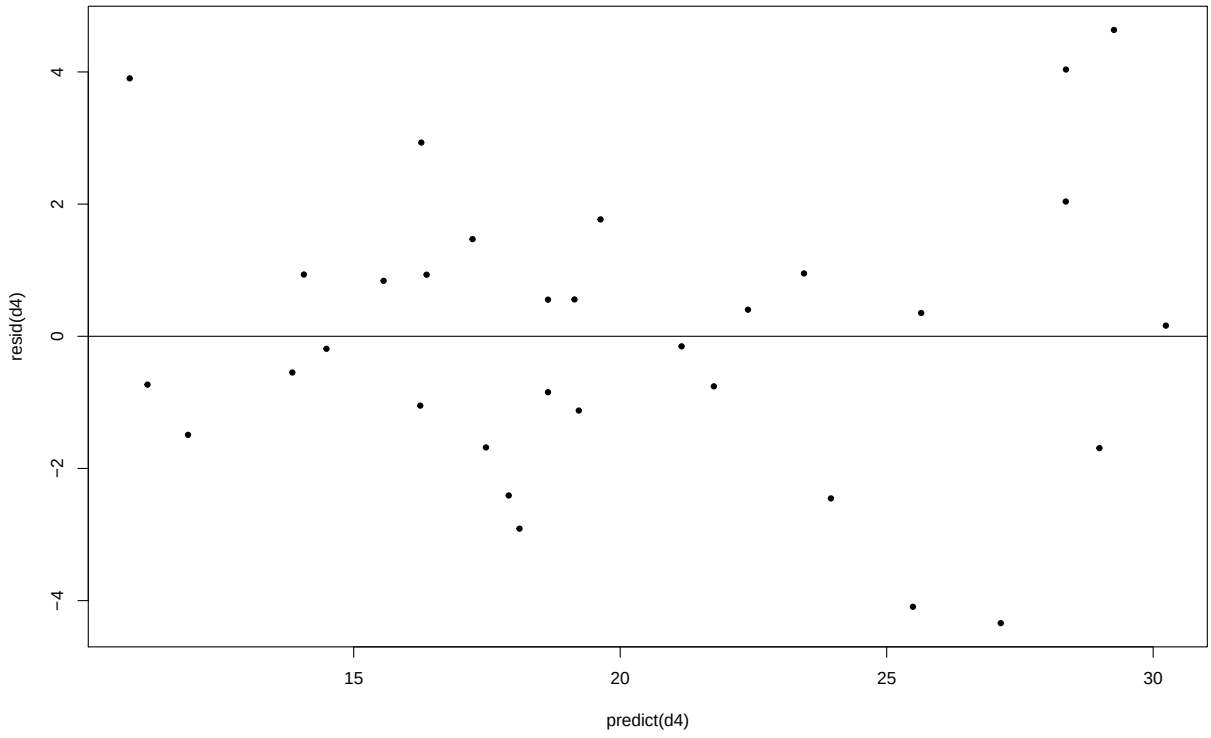
```
## Residual standard error: 2.397 on 25 degrees of freedom
## Multiple R-squared:  0.8724, Adjusted R-squared:  0.8418
## F-statistic: 28.49 on 6 and 25 DF,  p-value: 5.064e-10

## Predicted R Squared = 0.8051053

##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + vs, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3049 -1.1771 -0.1461  0.9042  5.3309
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.46561    2.38114  14.894 3.04e-14 ***
## cyl6        -3.15693    1.55442  -2.031 0.052603 .
## cyl8        -2.68035    2.69757  -0.994 0.329567
## hp          -0.02265    0.01224  -1.851 0.075605 .
## wt          -3.23718    0.75155  -4.307 0.000209 ***
## vs1          0.51612    1.58317   0.326 0.747030
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.482 on 26 degrees of freedom
## Multiple R-squared:  0.8578, Adjusted R-squared:  0.8305
## F-statistic: 31.37 on 5 and 26 DF,  p-value: 3.179e-10

## Predicted R Squared = 0.8016456

##  Note that the adj R squared & Predicted R squared start to degrade when we remove
##  am as a predictor

## Same happens if we remove cyl or vs

## Another approach is to remove wt as it is highly correlated to hp

##
## Call:
## lm(formula = mpg ~ cyl + hp + vs + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6694 -1.2032  0.4463  1.4205  4.7680
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.44721    2.57139   9.507 6.01e-10 ***
## cyl6        -2.65245    1.79591  -1.477  0.15170
## cyl8        -0.27710    3.48664  -0.079  0.93726
## hp          -0.04688    0.01451  -3.230  0.00335 **
```

```
## vs1          2.56903    1.94243   1.323  0.19749
## am1          5.16287    1.45386   3.551  0.00149 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.666 on 26 degrees of freedom
## Multiple R-squared:  0.8359, Adjusted R-squared:  0.8044
## F-statistic: 26.49 on 5 and 26 DF,  p-value: 1.973e-09
```

```
## Predicted R Squared = 0.7775341
```

```
## While we get a significant model with am as a significant predictor both Adjusted &
##   predicted R Squared degrade
```

```
## RESIDUALS PLOT
```




```
##                   2.5 %      97.5 %
## (Intercept) 24.14094312 38.228284595
## cyl6        -5.44443708  1.264219776
## cyl8        -6.18153296  6.763483782
## hp          -0.06321051 -0.006289991
## wt          -4.20147771 -0.545256467
## vs1         -1.63516397  5.615172015
## am1         -0.58833044  5.996019265
```