MotorTrend's Analysis of the Effects of Car Characteristics on MPG

S. Duffy

April 14, 2016

Summary

While the relationship between miles per gallon (MPG) and the type of transmission is significant, i.e. you're better off with a manual transmission, there are two other factors to consider as well. The weight of the vehicle and it's speed in the quarter mile are also significant indicators of miles per gallon.

Overview

Historically, the anecdote most car drivers pass down to their children is that a vehicle with a standard transmission is more fuel efficient than a vehicle with an automatic transmission. We here at MotorTrend Magazine love and trust our dads but we wanted to see if this anecdote stands up to an analysis of the data.

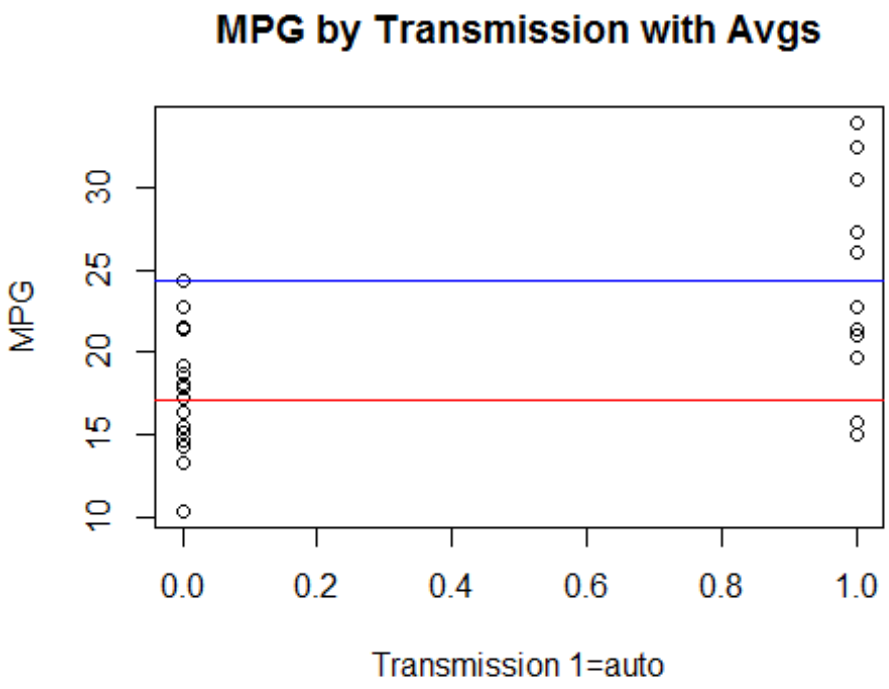
We were able to obtain rare data from industry insiders and have stored it in a dataset called mtcars. The definitions of the columns are thus:

- mpg = Miles/(US) gallon
- cyl = Number of cylinders
- disp = Displacement (cu.in.)
- hp = Gross horsepower
- drat = Rear axle ratio
- wt = Weight (1000 lbs)
- qsec = 1/4 mile time
- vs = V-style/Straight engine
- am = Transmission (0 = automatic, 1 = manual)
- gear = Number of forward gears
- carb = Number of carburetors

What we're interested in is the mpg and am columns which have the following characteristics.

The mpg column has a range from 10.4 mpg to 33.9 mpg with a mean of 20.09. Out of the 32 vehicles in the dataset, 13 are manual transmissions and 19 are automatic transmissions.

Is there are noticeable difference in MPG in the two different transmission types? A plot shows yes. The blue line is the average for 1s, the manual transmissions and the red line is the average for 0s, the automatic transmissions.



Is an automatic or manual transmission better for MPG?

For this section, we'll use a t-test for significance

```
ttest <- t.test(mancars$mpg,autocars$mpg,conf.level = .95)
```

In this case, the p-value is 0.0013736, which is lower than 0.05 so the null hypothesis that there is no difference in the MPG between transmissions is rejected. Also, the confidence interval does not contain zero, supporting the claim. Thus, a manual transmission is better for MPG.

Quantify the MPG difference between automatic and manual transmissions

But what about other models to fit the data and what impact do the variables have? One could argue all variables in the dataset could affect MPG. The model shows no variable is significant with our scattershot approach to determining MPG, although there is a relatively high R^2 at .869 (see Appendix). This could be overfitted, let's narrow it down and see if we get something better.

Based on conversations of my youth, the weight of a vehicle certainly has an impact and I know my old V8 had a much lower MPG than my old four cylinder. Also, I recall hearing that higher gears at highway speeds helps with MPG as well (model in Appendix)

The anova p-value is more significant than the model with all variables but two of our four variables are not significant, including transmission. Reviewing online literature, there is a step() function in the stats package that will discover which variables are significant. We'll input the model with all variables.

This shows that a model with transmission, weight, and speed in the quarter mile are significant. I would not have figured quarter mile speed to be significant. The residuals also show no correlation. This is the best model.

Conclusion

The optimal model uses transmission, weight, and speed in the quarter mile to predict miles per gallon. For every 1,000 pound increase in vehicle weight, MPG will decrease by nearly four miles. A manual transmission will net you nearly three miles per gallon and for every second longer to finish the quarter mile, you can expect an additional 1.22 miles per gallon.

Appendix

First comparison, model using all variables.

```
mdl <- lm(mpg ~ am,mtcars)
mdltotal <- lm(mpg ~ ., mtcars)
summary(mdltotal)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

The model based on guesses from my youth.

```
mdl2 <- lm(mpg ~ am + wt + cyl + gear, mtcars)
summary(mdl2)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + cyl + gear, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.529 -1.491 -0.524  1.494  5.622
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  43.1678     4.5197   9.551 3.77e-10 ***
## am            1.3827     1.7583   0.786  0.43851
## wt           -3.0911     0.9108  -3.394  0.00214 **
## cyl          -1.5688     0.4258  -3.684  0.00101 **
## gear         -1.0812     1.0579  -1.022  0.31584
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.61 on 27 degrees of freedom
## Multiple R-squared:  0.8367, Adjusted R-squared:  0.8125
## F-statistic: 34.57 on 4 and 27 DF,  p-value: 2.928e-10
```

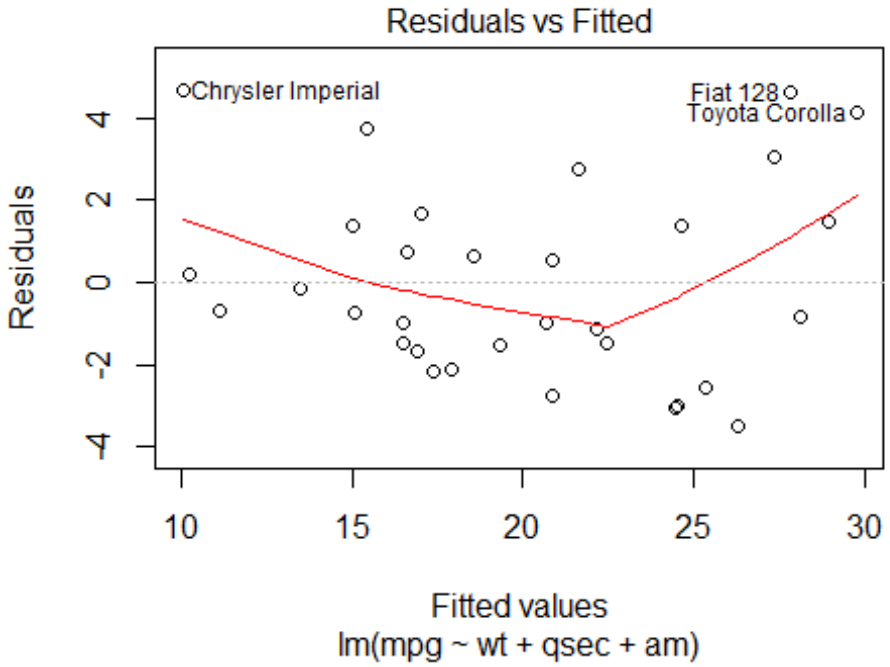
The model based on the step() function.

```
steps <- step(mdltotal, trace=0)
summary(steps)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am            2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

Plot of the residuals from the step function:

```
plot(steps)[1]
```



!C:\Users\stephen.p.duffy\Documents\GitHub\LinearRegression\MPG_Analysis_files\figure-docx\stepplot-2.png!C:\Users\stephen.p.duffy\Documents\GitHub\LinearRegression\MPG_Analysis_files\figure-docx\stepplot-3.png!C:\Users\stephen.p.duffy\Documents\GitHub\LinearRegression\MPG_Analysis_files\figure-docx\stepplot-4.png

```
## NULL
```