Are automatic or manual transmissions better for MPG

BK

Thursday, October 22, 2015

Executive Summary

We cannot determine if Automatic or Manual transmissions are better. We arrive at this conclusion by building a rigorous model with relevant predictors and validating the model with statistical measures.

Exploratory Analysis

We start by probing the data provided and understanding all the factors provided, using the ?mtcars command. This information is critical to model selection as seen below.
Then, we generate a plot of transmission (am) v/s MPG, in the Appendix, that shows that there is a clear improvement in MPG from automatic (am=0) to manual (am=1) transmissions. But notice the wide dispersion in data points. Clearly, transmission alone does not adequately explain the situation. So we begin our detailed analysis.

Choosing a model

We choose a model by asking several questions till we arrive at a satisfactory answer.

Does transmission alone explain the change in MPG?

We find that it does not. By running the commands "fit <- lm(mpg~as.factor(am), data = mtcars)" and "summary(fit)$coef" we get the false comfort that manual transmission improves MPG by 7.245. Even the t value and p statistics of 4.106 and 2.8510^{-4} respectively, implying that the null hypothesis should be rejected are misleading (null hypothesis is that there is no difference due to transmission type). We also see from the adjusted R squared value of 0.338, that a rather low proportion of the variability is explained.

Does including all the given factors (10) as regressors explain the change in MPG(11)?

Doing so proves the 'garbage in garbage out' thumb rule. No meaningful conclusion can be arrived at. The large values for p-statistics for the 10 regressors make it clear that the null hypothesis should not be rejected. Specifically for transmission, the p-statistic is high at 0.234 compared to the typical 0.05.

What contextual information can be applied to improve our analysis?

We know that adding redundant regressors increases the standard error of valid regressors. So, let's examine which factors are valid and which are redundant regressors,

1. Gross horsepower (hp): From common sense/ context, we know this is valid regressor. Now, compared to Gross horsepower, which of the given factors are redundant? The Number of cylinders (cyl), Displacement (disp), 1/4 mile time (qsec) are all directly or inversely proportional to Gross horsepower. So they provide no new insight and hence are redundant. The Number of carburetors (carb) is typically a function of Number of cylinders (cyl) and hence not relevant by itself.
2. Weight (lb/1000) (wt): Again, from context, we know this is an important predictor of MPG.
3. Transmission (am): Transmission is the regressor of interest and needs to be included in the analysis. The Rear axle ratio (drat) and the Number of forward gears (gear) are elements of the transmission and are not relevant by themselves.
4. Other redundant factors: There is no explanation for V/S (v/s) and is dropped.

So, we're left with the following as relevant regressors: Gross horsepower (hp), Weight (lb/1000) (wt) and Transmission (am).

Model selection through nesting

Lets' start by fitting MPG against Transmission alone (recap from earlier). We found that manual transmission improves MPG by 7.245. The t value and p statistics were 4.106 and 2.8510^{-4} respectively. But the adjusted R squared value of 0.338 made it clear that not enough of the variability was explained.

So, let's expand our regressors by fitting MPG against Transmission and Gross horsepower. We see the impact of a manual transmission is an improvement to MPG by 5.277. The t value and p statistics were 4.888 and 3.510^{-5} respectively. The adjusted R squared value of 0.767 shows a substantial improvement.

Lastly, let's add Weight (lb/1000) (wt) to the regressors by fitting MPG against Transmission, Gross horsepower and Weight (lb/1000).

The full set of coefficients shows regressor slopes/ coefficients that are consistent with what we know in the real world - MPG reduces with weight and horsepower.

```
##                   Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)    34.00287512 2.642659337 12.866916 2.824030e-13
## as.factor(am)1  2.08371013 1.376420152  1.513862 1.412682e-01
## hp             -0.03747873 0.009605422 -3.901830 5.464023e-04
## wt             -2.87857541 0.904970538 -3.180850 3.574031e-03
```

The adjusted R squared value of 0.823 shows a very high proportion of variability being explained.

Testing our model and the inferences we can make from it

Confidence Interval test

As seen below, while horsepower and weight clearly show a statistically significant relationship with MPG, transmission type does not. Zero is included in the confidence interval for transmission type.

```
##                      2.5 %      97.5 %
## (Intercept)    28.58963286 39.41611738
## as.factor(am)1 -0.73575874  4.90317900
## hp             -0.05715454 -0.01780291
## wt             -4.73232353 -1.02482730
```

Residuals, influence, leverage and outlying points test

The charts in the Appendix prove that,
1. In the Residuals v/s Fitted plot, the points are distributed above and below zero and there is no systematic pattern to indicate poor model fit. This is as desired.
2. The Residual QQ plot shows the normality of errors, again, as desired.

ANOVA test

The results of the ANOVA test, in the Appendix, shows the validity of adding relevant regressors through the improving RSS (residual sums of squares), the F statistic and p statistic.

Conclusion

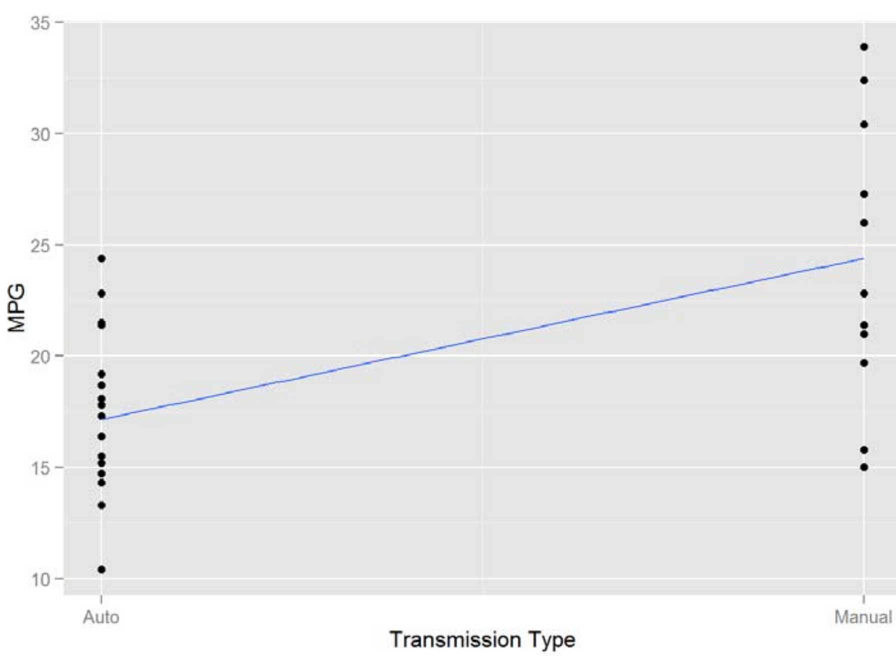
Thus, we believe that we've arrived at a good model for MPG and its key regressors.

However, we are unable to answer the questions asked about which type of transmission is better and by how much because the confidence interval spans zero.

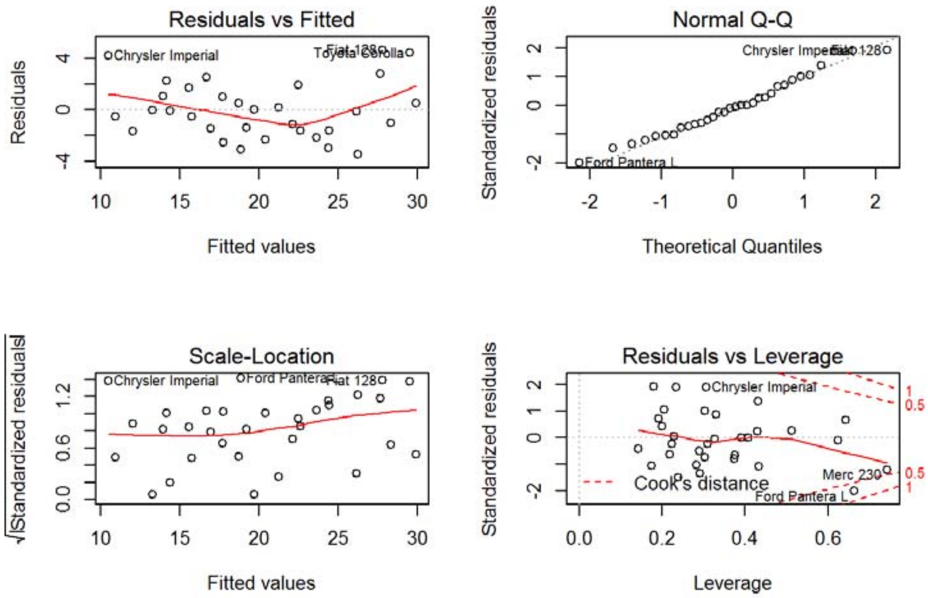
Appendix

Tranmission Type v/s MPG

```
## Warning: package 'ggplot2' was built under R version 3.1.3
```



Residual plots



ANOVA test for model selection

```
## Analysis of Variance Table
##
## Model 1: mpg ~ as.factor(am)
## Model 2: mpg ~ as.factor(am) + hp
## Model 3: mpg ~ as.factor(am) + hp + wt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 245.44  1    475.46 73.841 2.445e-09 ***
## 3     28 180.29  1     65.15 10.118  0.003574 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```