Automatic or Manual - Which is Best for MPG?

Andrew Holland

11/12/2020

Executive Summary

Using the "mtcars" dataset, we look to address the following questions:

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

From our investigation, we conlcude that a manual car typically has better MPG, although this is not directly due to causation - manual cars are typically lighter and less powerfulWe also conlcude that there is no significant impact on MPG caused by the type of transmission the car has.

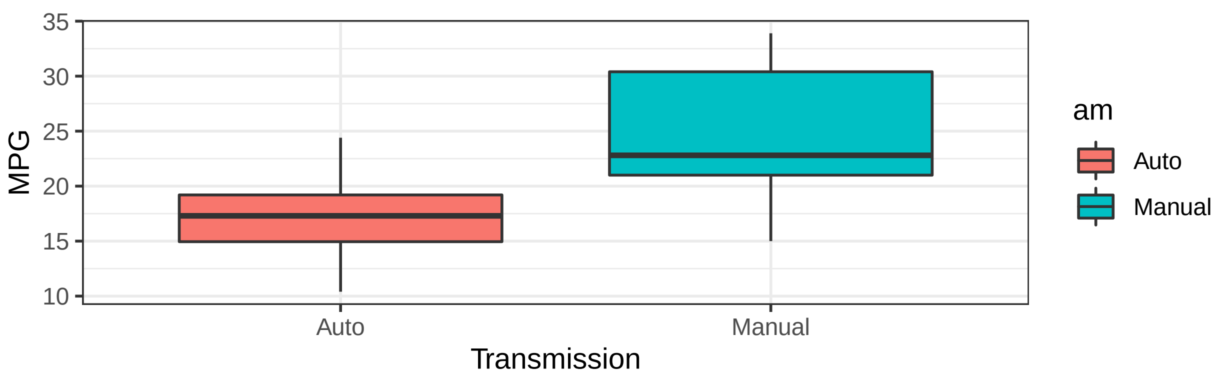
It is worth noting however that the dataset used may not be the entire picture - as suggested from our diagnostic plot, we could claim that there are "known unknowns" influencing our model, that we have not accounted for in our final model. What we do know is that transmission type is not this unaacounted-for variable.

Initial Exploration

As a first step, lets look at the dataset:

```
##                    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
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
```

For each car model, we have 10 measured values. Intially, the most important measurements to us are MPG and am (representing automatic and manual as a 0 or 1 respectively). Lets plot these, and see if we find a relation:



From this brief assessment, we could naturally conclude that "manual cars have a better MPG than automatic cars." We can also peform a t-test to verify that this conclusion is valid (Appendix 1). This however is not the whole picture - we have shown a trend, but not determined causation. It could be the case that the transmission has no impact on the MPG, but some other variable instead.

Lets assume that the mtcars dataset contains all the factors determining a car's MPG We then have 9 other measurements besides transmission that could influence a car's MPG Introducing these into our model will begin to pull back the curtain on how inflential trnasmission truly is on MPG.

Applying Models

We could simply create a linear model that predicts mpg using all the other variables in the dataset (Appendix 2). As the number of variables (9) is large compared to the number or observations (32), the model suffers the consequences of overfitting - our R-squared and p-values are reasonable, but the model terms are largely insignificant. Here we have begun to model the random error in the data, rather than the underlying relationships between our variables.

We can however look a the significance values, and we see that two variables, weight (wt) and horesepower (hp) are both below 0.1 - far more significant than the rest (the next closest is the intercept at 0.25.) We will therefore start by investigating the impact of these two variables (along with transmission type) on mpg.

Starting with weight, and then adding horsepower as variables, we find a reasonable model, with high significace. We should also consider that there is also some underlying relation between weight and horsepower (not an unreasonable assumption), which we add in model 3. Lastly, we introduce our study variable for transmission type am, producing model 4. We can compare these models using the anova() function.

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt
## Model 2: mpg ~ wt + hp
## Model 3: mpg ~ wt + hp + wt:hp
## Model 4: mpg ~ wt + hp + wt:hp + am
##   Res.Df    RSS Df Sum of Sq       F   Pr(>F)
## 1     30 278.32
## 2     29 195.05  1    83.274 17.3328 0.000287 ***
## 3     28 129.76  1    65.286 13.5888 0.001009 **
## 4     27 129.72  1     0.042  0.0088 0.925942
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here, we see that introducing the weight/horespower interaction has a significant effect on the model (model 3), but that introducing the effect of transmission type (model 4) produces very little impact, with a p-value greater than 0.9.

Looking at these models, we would select model 3 as the best - a model where transmission is not included.

Before making conclusions, let's investigate the diagnostic plots of the model (Appendix 3).

- From the Residuals vs Fitted plot, we can see that there is no clear trend in the residuals, with the average line staying close to zero.
- The Normal Q-Q plot tails off at the lower quantiles, but otherwise follows a roughly straight line, suggesting that the values observed in the actual mtcars data could feasibly have come from our model.
- In the Scale-Location plot shows a curious upward trend on the latter half of the fitted values. This could be caused by an variable that we haven't accounted for.
- The Residuals vs Leverage plot is promising, with no points at or beyond the Cook's Distance, and so no one point is overly influencing our model.

Conclusions

From our model 3, we can draw the following conclusions:

- Increasing weight by 1 unit (1000 lbs) will result in an expected decrease of 8.21 MPG
- Increasing horsepower by 10 units (gross hp) will result in an expected decrease of 0.12 MPG
- The above values are average decreases - there is an interaction between them - the actual decrease per unit of one variable is impacted by the other variable.
- There is no significant impact on MPG caused by the type of transmission the car has.

Appendix 1

```
data_man <- mtcars[mtcars$am == 1,]
data_auto <- mtcars[mtcars$am == 0,]
ttest <- t.test(data_auto$mpg, data_man$mpg, alternative = "two.sided")
ttest
```

```
##
##  Welch Two Sample t-test
##
## data:  data_auto$mpg and data_man$mpg
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean of x mean of y
##  17.14737  24.39231
```

Appendix 2

```
summary(overmodel)
```

```
##
## Call:
## lm(formula = mpg ~ factor(cyl) + disp + hp + drat + wt + qsec +
##     factor(vs) + factor(am) + factor(gear) + factor(carb), 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
## factor(cyl)6   -2.64870    3.04089  -0.871   0.3975
## factor(cyl)8   -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
## factor(vs)1     1.93085    2.87126   0.672   0.5115
## factor(am)1     1.21212    3.21355   0.377   0.7113
## factor(gear)4   1.11435    3.79952   0.293   0.7733
## factor(gear)5   2.52840    3.73636   0.677   0.5089
## factor(carb)2  -0.97935    2.31797  -0.423   0.6787
## factor(carb)3   2.99964    4.29355   0.699   0.4955
## factor(carb)4   1.09142    4.44962   0.245   0.8096
## factor(carb)6   4.47757    6.38406   0.701   0.4938
## factor(carb)8   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
```

Appendix 3

```
par(mfrow =c(2, 2))
plot(mdl3, labels.id=mtcars$names)
```