Motor Trend Investigates : Was an automatic or manual transmission better for MPG for 1973-1974 cars?

Pasi Hyytiäinen

May, 2015

Summary

The data was extracted from the old 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). We are interested to answer two question: "Is an automatic or manual transmission better for Mpg?" and "How different is the Mpg between automatic and manual transmissions?"

The first operation, a t-test, will show that a performance difference existed between both types of cars. As the dataset is quite small, it can't be splitted to more meaningful groups for example based on weight and cylinder count. Even with this small dataset, we found out correlationship between the various significant variables and the MPG.The end result of the current evaluation showed that Manual cars to have better MPG indices , making them the more efficient cars. For each unit of weight (1 ton) increase, manual cars had a 2.93 increase in MPG, covring more miles than with an Automatic car. Fuel efficency point of view light manual car is better than heavy automatic car.

R code, can be found from [git] (https://github.com/LilaLipetti/MtCarsAnalysis)

Data analysis

In appendix 1, there is preliminary analysis of the data. Based on the Appendix 1, Figure 1, it seems like there might be some diffrences between Automatic And Manual cars, so we need first test if **H0** = *There is no performance difference (MPG) between Automatic and Manual cars*

Based on the physics, we know that more heavier the car is, more energy is needed to move the car. That would mean that heavier car needs more hp to move it as smoothly as smaller car. While analysing the data we can see based on appendix 1, table 1 that our cars with automatic transmission are much heavier than our manual transmission cars. So we can say that transmission type is not normally distrubuted based on weight.

As the motor is one of the heavier part of a car, we could try to group our data based on cylinders and transmission type. As we can see in appendix 1, table 2: 12 of our 32 cars are 8 cylinder, has automatic transmission and it's the heaviest car group with lowest mean mpg. Although 2nd biggest group has 8 cars with 4 cylinder, manual transmission with biggest mean mpg.

Note! The most optimal data should contain same amount of cars in each cylinder/transmission type group so that weight and hp is normal distributed based on transmission type. Also another significant improvement for the dataset would be that it contains data for same carmodel but with manual and automatic transmission type. In current dataset we can't find such car.

Even though that the dataset isn't optimal, we could try to determine the correlations between the variables in the dataset. We make scatterplot matrix using pairs function(Appendix 2 Plot 1). From that plot, we can see that many variables like cyl, disp, hp, drat, wt, vs, am and carb have correlations with mpg. Those correlations needs to be considered while finding the best linear model.

Hypothesis Testing (t-test):

The null hypothesis was formulated and tested for possible rejection:

H0 = *There is no performance difference (MPG) between Automatic and Manual cars*

As the **t-test** compares the means of two data sets, `mtcars` data was split into **"automatic"** and a **"manual"** data subsets. The **r code** for the test is `t.test(data_aut,data_man)`. Results obtained were (t = -3.7671231), (df = 18.3322516), (p-value = 0.0013736). This P-value of 0.0013736 allows us to confidently reject the null hypothesis and accept the alternate hypothesis **H1**:

H1 = *There exists a performance difference between Automatic and Manual cars*

Regression Analysis

We need to build linear regression models using different variables in order to find the best fit. The `step` function in **r** can be used for this operation and produces the best fit model. A forward elimination and backward tests were used to validate the previous result.

Full elimimination gave:
lm(formula = mpg ~ wt + qsec + am, data = cars)
Multiple R-squared: 0.8497, Adjusted R-squared: 0.8336
F-statistic: 52.75 on 3 and 28 DF, p-value: 1.21e-11

Forward elimination gave:
lm(formula = mpg ~ am + hp + wt + qsec, data = cars)
Multiple R-squared: 0.8579, Adjusted R-squared: 0.8368
F-statistic: 40.74 on 4 and 27 DF, p-value: 4.589e-11

Backward elimination gave: lm(formula = mpg ~ wt + qsec + am, data = cars)
F-statistic: 52.75 on 3 and 28 DF, p-value: 1.21e-11
Multiple R-squared: 0.8497, Adjusted R-squared: 0.8336

The forward elimination introduced **hp** into the model. Based on physics both models might not be optimal as very light car with high hp will get very good 1/4 mile time values (qsec) compared to heavy, low hp cars. Even if the p-value: 1.21e-11 of the backward elimination is better than for forward elimination, the forward elimination model explains mpg a little bit better 85,79% > 84,97%. So the *lm(formula = mpg ~ am + hp + wt + qsec, data = cars)* will be analyzed more.

Residuals and diagonistics

To check residuals for normaility and homoskedasticity, we plot the residuals(see Appendix 5). We find that the residuals are normally distributed and homoskedastic for *lm(formula = mpg ~ am + hp + wt + qsec, data = cars)*

Results

Model Accuracy

Based on the confidence intervals (see Appendix 4), selected model *lm(formula = mpg ~ am + hp + wt + qsec, data = cars)* will be considered to be significant.

Interpert Results

We rejected **H0** = *There is no performance difference (MPG) between Automatic and Manual cars* because the P-value of 0.0013736 allows us to confidently reject the null hypothesis. The mean for manual transmission is *24.3923 Mpg* and *17.1474 Mpg* for automatic transmission, which is a significant difference in the means. Manual transmissions have a higher value so based on mean of mpg it is better (cheaper) to have manual transmission.

```
##                Estimate Std. Error   t value    Pr(>|t|)
## (Intercept) 17.44019110  9.3188688  1.871492 0.072149342
## ammanual     2.92550394  1.3971471  2.093913 0.045790788
## hp          -0.01764654  0.0141506 -1.247052 0.223087932
## wt          -3.23809682  0.8898986 -3.638726 0.001141407
## qsec         0.81060254  0.4388703  1.847021 0.075731202
```

Intercept The intercept is at 17.4 mpg.

mpg vs hp Negative slope meaning that when hp increases, miles per Gallon decreases by a factor of 0.01 Miles less Per hp (-0.01) as horspower increases.

mpg vs weight Negative slope meaning that when weight increases, the MPG index goes down.
3.3 miles less (-3.3) per Gallon for every increase of 1000lb in the weight.

mpg vs qsec (1/4 mile time) Positive slope, meaning that when the "Quarter Mile Time Index" increases by one unit (car is slower), the car becomes more efficient by covering an additional 0.81 Mile per Gallon.

am=2.93 Positive slope. A manual car has an MPG index of 2.93. Meaning it will cover 2.93 more miles compared to an automatic car, making it a more efficient vehicle.

Final Conclusion

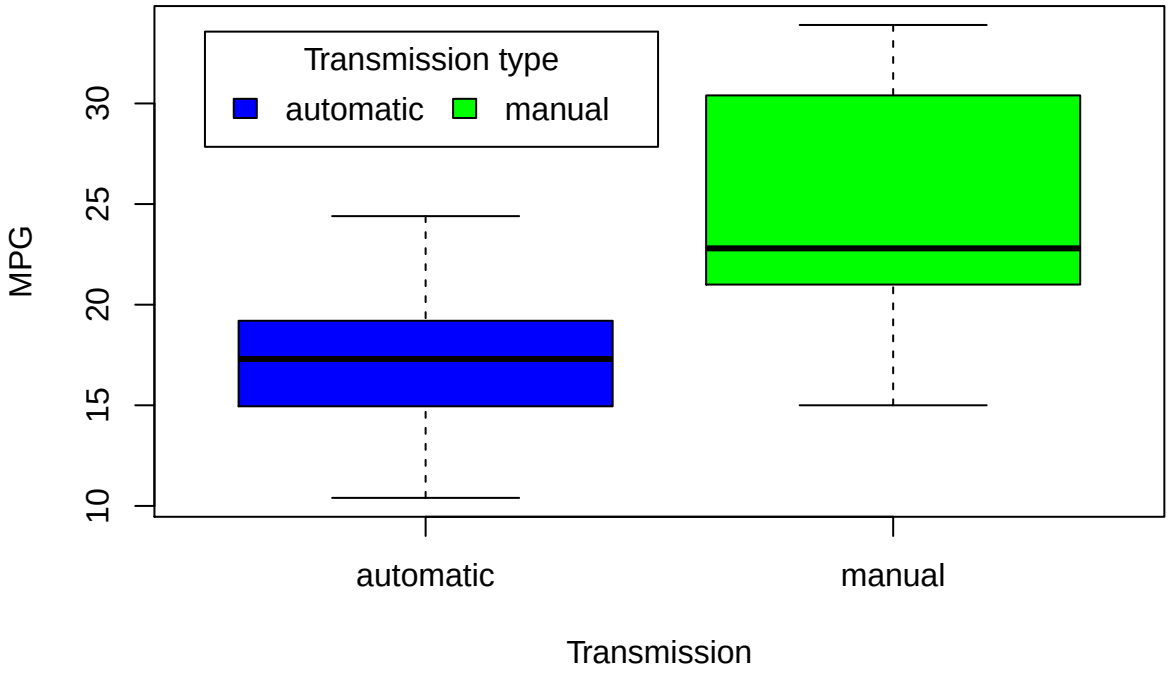
H1 = *There exists a performance difference between Automatic and Manual cars*

Based on the selected model *lm(formula = mpg ~ am + hp + wt + qsec, data = cars)*, manual cars are more efficient and, for every unit (1 ton) of car weight increase, manual cars have an extra range of 2.93 Miles per gallon compared to automatic cars.

Appendix

Appendix 1 : `mtcars` Data Exploration

Appendix 1, figure 1: Boxplot of transmission against MPG



Let's investigate the data more detailed. We will group data based on the transmission and then based on the transmission and the cylinder count.

Appendix 1, table 1 : Data summary based on transmission type

```
##          am carcount  mean_hp    sd_hp mean_mpg   sd_mpg  mean_wt
## 1 automatic       19 160.2632 53.90820 17.14737 3.833966 3.768895
## 2    manual       13 126.8462 84.06232 24.39231 6.166504 2.411000
##       sd_wt
## 1 0.7774001
## 2 0.6169816
```

Appendix 1, table 2 : Data summary based on cylinder count and transmission type

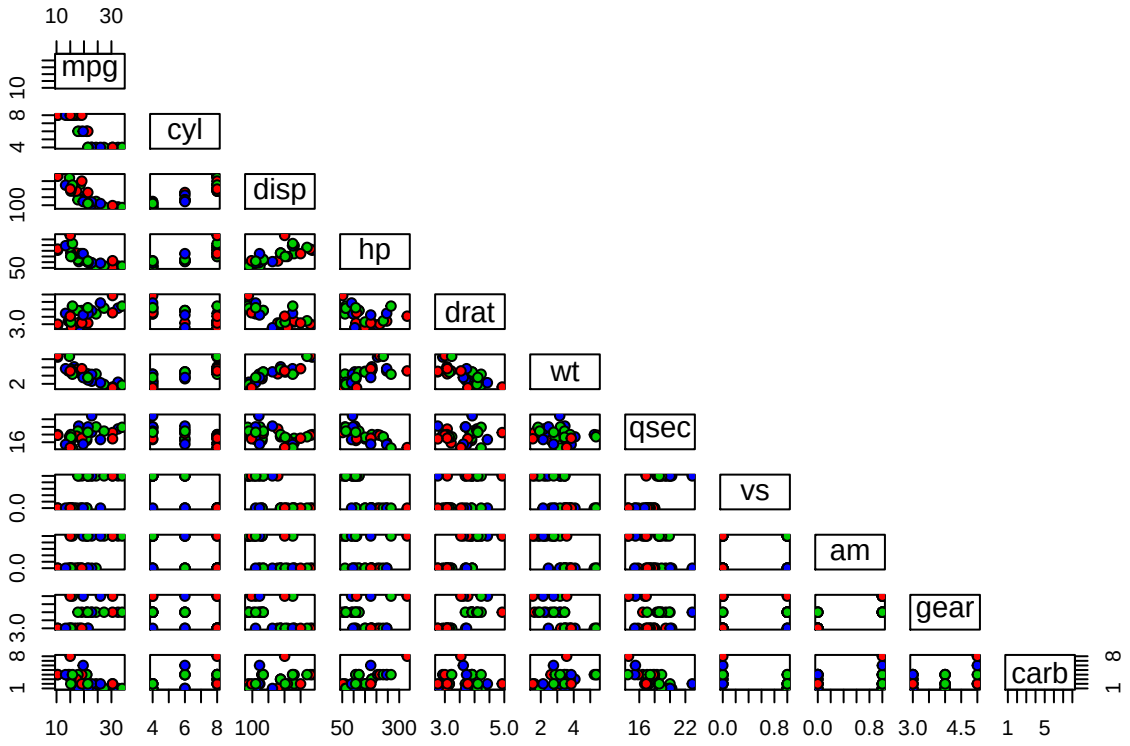
```
##   cyl        am carcount   mean_hp    sd_hp mean_mpg    sd_mpg  mean_wt
## 1   4 automatic        3  84.66667 19.65536 22.90000 1.4525839 2.935000
## 2   4    manual        8  81.87500 22.65542 28.07500 4.4838599 2.042250
## 3   6 automatic        4 115.25000  9.17878 19.12500 1.6317169 3.388750
## 4   6    manual        3 131.66667 37.52777 20.56667 0.7505553 2.755000
## 5   8 automatic       12 194.16667 33.35984 15.05000 2.7743959 4.104083
## 6   8    manual        2 299.50000 50.20458 15.40000 0.5656854 3.370000
```

```
##       sd_wt
## 1 0.4075230
## 2 0.4093485
## 3 0.1162164
## 4 0.1281601
## 5 0.7683069
## 6 0.2828427
```

Appendix 2 : Scatter plot for `mtcars`

```
pairs(mtcars,pch = 21, bg = c("red", "green3", "blue"),upper.panel = NULL)

title(sub="Appendix 2 Plot 1: Mtcars Data - paired scatter plot")
```



Appendix 2 Plot 1: Mtcars Data − paired scatter plot

Appendix 3 : Statistics for Models

Forward elimination gave:

```
##                Estimate Std. Error   t value    Pr(>|t|)
## (Intercept) 17.44019110  9.3188688  1.871492 0.072149342
## ammanual     2.92550394  1.3971471  2.093913 0.045790788
## hp          -0.01764654  0.0141506 -1.247052 0.223087932
## wt          -3.23809682  0.8898986 -3.638726 0.001141407
## qsec         0.81060254  0.4388703  1.847021 0.075731202
```

Backward elimination gave:

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## ammanual     2.935837  1.4109045  2.080819 4.671551e-02
```

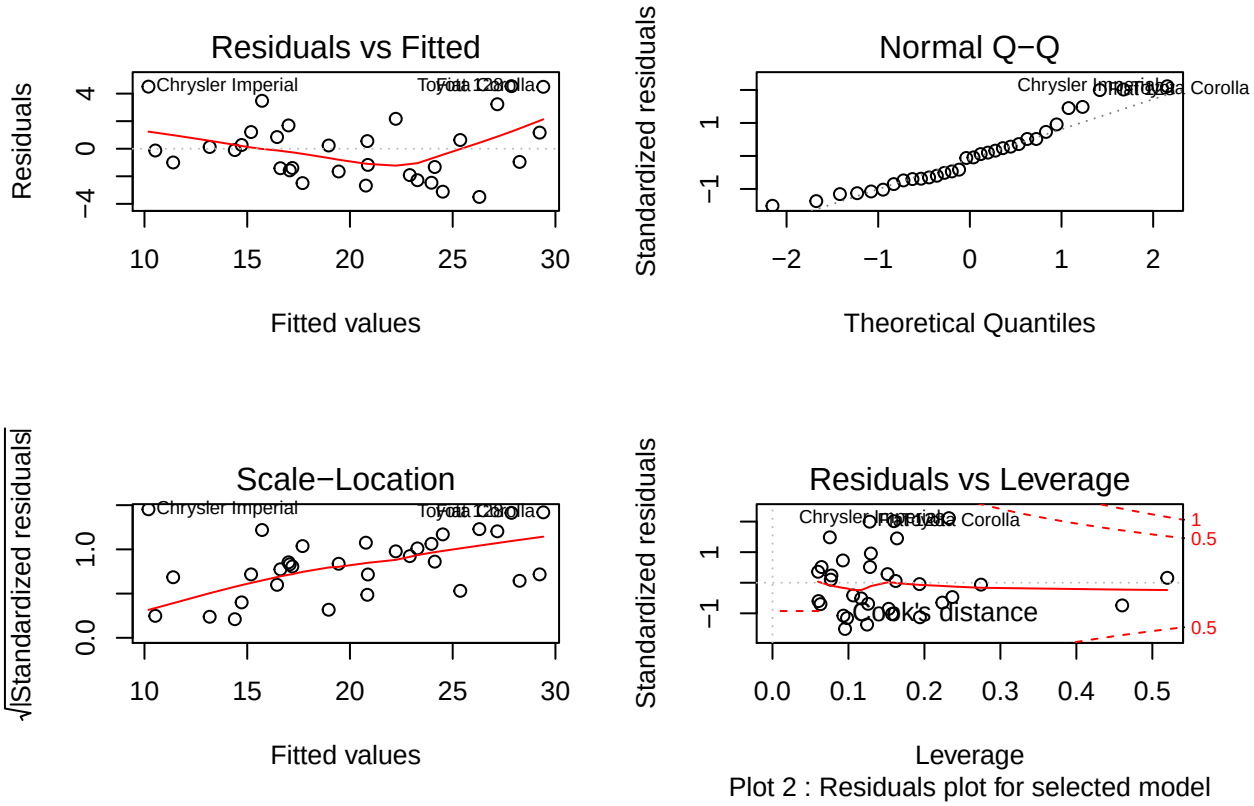
Anova comparision

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + hp + wt + qsec
## Model 3: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     27 160.07  3    560.83 26.6167 2.382e-07 ***
## 3     21 147.49  6     12.57  0.2983    0.9308
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix 4 Confidence interval

```
##                   2.5 %     97.5 %
## (Intercept) -1.68054821 36.5609304
## ammanual     0.05879488  5.7922130
## hp          -0.04668118  0.0113881
## wt          -5.06401789 -1.4121758
## qsec        -0.08988486  1.7110899
```

Appendix 5 : Residual Plots



Plot 2 : Residuals plot for selected model