Regression Model Project, MPG vs. Transmission

Jerin Timothy James

2023-09-09

Executive Summary

In the role as a Motor Trend employee (a magazine about the automobile industry) the following, the below analysis of the mtcars data set is presented to answer questions about Miles Per Gallon (mpg) differences. The two questions of particular interest are:

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

Answers:

1 & 2. Taken as a solo variable, a manual transmission is approx 7 mpg more efficient.

Note: When looked at in a larger context, a better mpg model can be put together with variables such as weight, engine size, and gearing included.

Exploratory Data Analysis

First we need to load the required data set and examine its structure with the below code:

```
#Load needed library and data
library(ggplot2)

#Load data
data(mtcars)
#?mtcars

#according to ?mtcars, The data was extracted from the 1974 Motor Trend US magazine,
#and comprises fuel consumption and 10 aspects of automobile design and
#performance for 32 automobiles (1973-74 models). And that it is A data frame
#with 32 observations on 11 (numeric) variables.
#*Note that [, 9]   am  Transmission (0 = automatic, 1 = manual)

#Initial data exploration
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
```

```
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
head(mtcars)
```

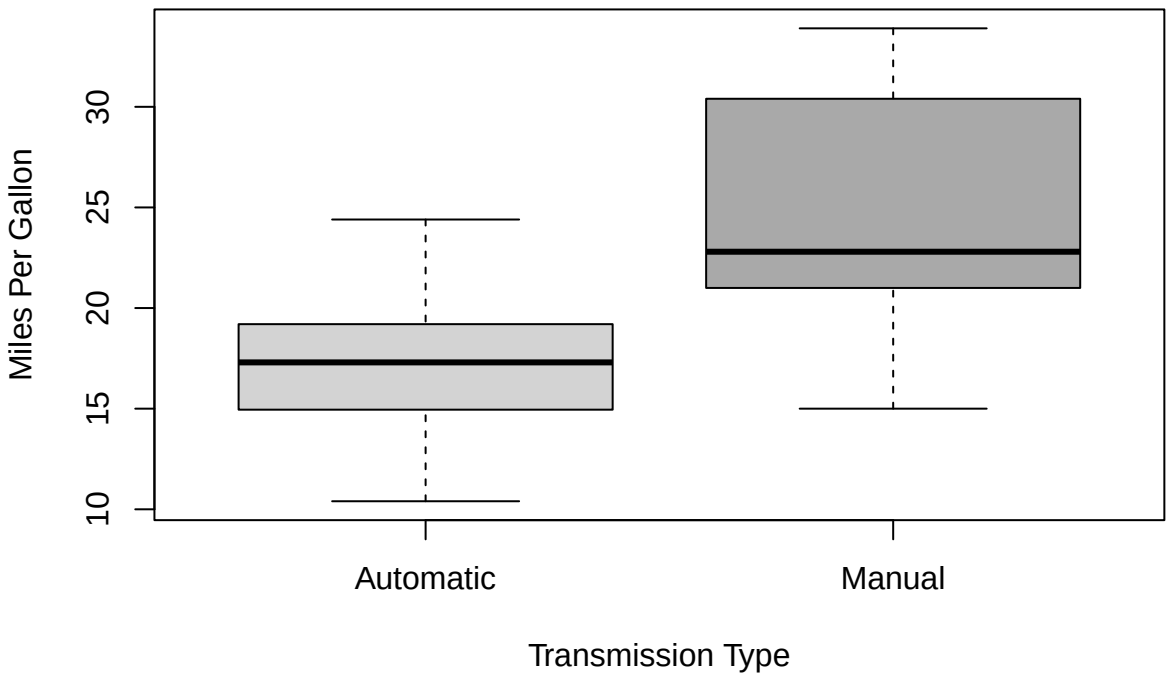
```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 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
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Next we classify some of the variables as factors to ease functions such as plotting with informative labels:

```
#Change to factors for ease
mtcars$vs <- factor(mtcars$vs)
mtcars$am <- factor(mtcars$am, labels=c("Automatic","Manual"))
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
mtcars$cyl <- factor(mtcars$cyl)
```

Now we can take a simple approach to see if the type of transmission affects mpg with the following code:

```
#Visual and mean validation of difference in mpg
boxplot(mpg ~ am, mtcars, col = (c("lightgrey","darkgrey")), ylab = "Miles Per Gallon",
        xlab = "Transmission Type")
```




```
aggregate(mtcars$mpg,by=list(mtcars$am),FUN=mean)
```

```
##     Group.1        x
## 1 Automatic 17.14737
## 2    Manual 24.39231
```

We see that there is a significant difference based on the provided data. The automatic transmissions are averaging 17.1 mpg while the manuals are averaging 24.4 mpg. The box plots further show that the quantiles of data points are not in significant overlap.

Linear Regression Provides More Detailed Analysis

With a data set more complete than just transmission and mpg data, we can investigate if that is a complete model, or if there may be more/additional influential factors in action.

The below code is a linear regression of transmission vs mpg and shows, via a low Rsquared value, that in the data collected, perhaps only 1/3 of the difference in mpg of the 32 cars is properly assigned to transmission alone.

```
#Linear regression for more in depth mpg validation
lmam<-lm(mpg~am,mtcars)
summary(lmam)
```

```
##
```

```
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## amManual       7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

The next step taken was to look at an analysis of variation amoung the various variables given in the mtcars data set to see if additional influencers could be easily found in the below code.

```
#Brief look for influential variables
summary(aov(mpg ~ ., mtcars))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## cyl          2  824.8   412.4  51.377 1.94e-07 ***
## disp         1   57.6    57.6   7.181   0.0171 *
## hp           1   18.5    18.5   2.305   0.1497
## drat         1   11.9    11.9   1.484   0.2419
## wt           1   55.8    55.8   6.950   0.0187 *
## qsec         1    1.5     1.5   0.190   0.6692
## vs           1    0.3     0.3   0.038   0.8488
## am           1   16.6    16.6   2.064   0.1714
## gear         2    5.0     2.5   0.313   0.7361
## carb         5   13.6     2.7   0.339   0.8814
## Residuals   15  120.4     8.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at this data, and noticing only 5 of the 11 variables are under a 0.25 p value, another model was proposed including all of those factors in the below code. (Note, horsepower is under 0.25 but not a physical trait difference, but a performance product of the other included variables, and was therefor not included).

```
#More inclusive model for mpg variance
lmam2<-lm(mpg ~ am + cyl + disp + wt + drat, mtcars)
summary(lmam2)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + disp + wt + drat, data = mtcars)
##
## Residuals:
```

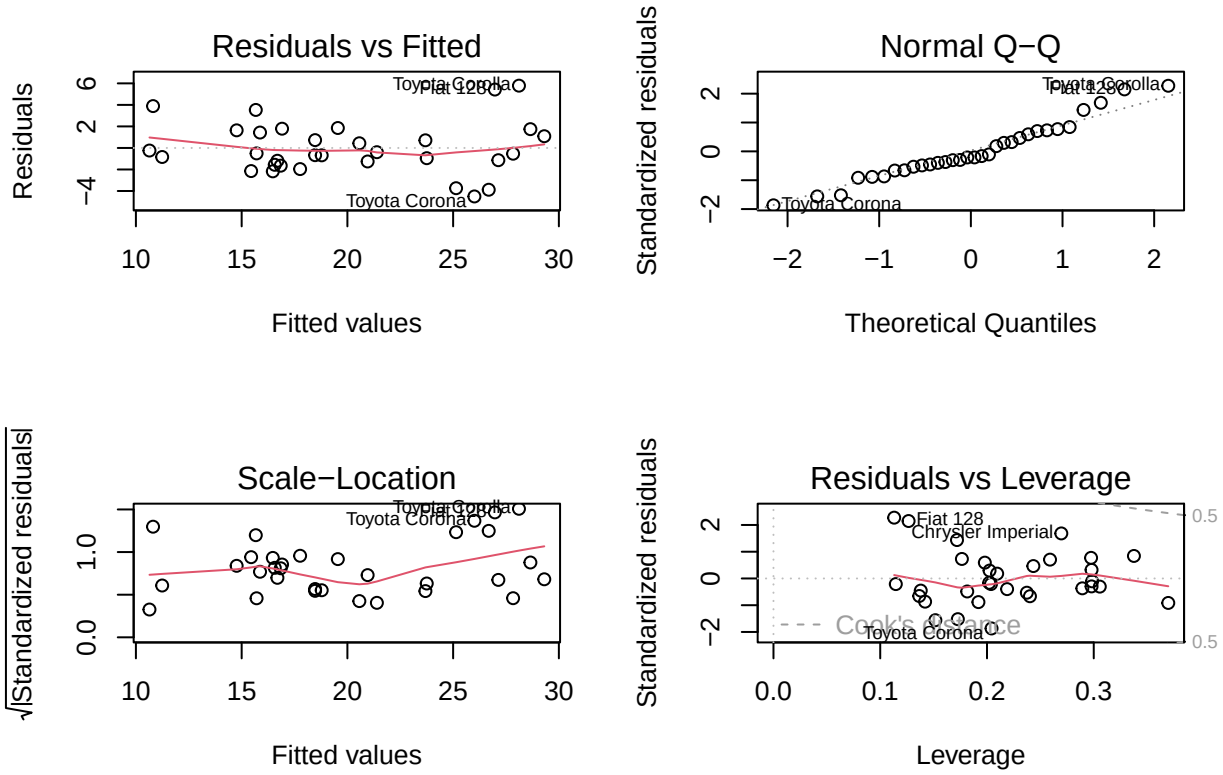
```
##     Min      1Q  Median      3Q     Max
## -4.5067 -1.3347 -0.5209  1.4828  5.7861
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.796707   6.702381   5.192 2.27e-05 ***
## amManual     0.261366   1.539697   0.170   0.8666
## cyl6        -4.374626   1.580145  -2.768   0.0105 *
## cyl8        -6.407638   2.753475  -2.327   0.0284 *
## disp         0.001425   0.014079   0.101   0.9202
## wt          -3.251684   1.273250  -2.554   0.0171 *
## drat        -0.255597   1.565998  -0.163   0.8717
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.703 on 25 degrees of freedom
## Multiple R-squared:  0.8378, Adjusted R-squared:  0.7988
## F-statistic: 21.52 on 6 and 25 DF,  p-value: 9.46e-09
```

Looking at the RSquared in this model shows much more (over 80%) of the mpg variation is explained. This model fits with a commmon sense understanding of mpg influencers too (weight, engine size, gearing)

Model Validation

Finally we take a quick look at the residual plots below to show that they meet the criteria to consider the model unbiased, namely that the residuals are homoscedastic and normally distributed, with the exception of a few outliers.

```
#Residual analysis for model validation
par(mfrow = c(2, 2))
plot(lmam2)
```



Conclusion:

The data supports the claim that manual transmissions are more fuel efficient than automatic transmissions. The data also supports there are other significant factors that change mpg, such as weight, engine size, and gear ratios. Further data gathering and analysis is recommended for a more precise model as this data set was very small in both size and variety relative to the quantity of vehicles in operation and the number of vehicle compositions available.