Regression Modeling Course Project

Peter Sedivec

June 20, 2015

Executive Summary

The objective of this analysis is to answer the following two questions: (1) Is an automatic or manual transmission better for MPG and (2) can we quantify the MPG difference between the two transmission types. We would like to publish an article in Motor Trend answering these two questions and informing our readers so they can be educated buyers in making their next automobile purchase.

The mtcars dataset was used to perform this analysis which has 32 observations. The analysis demonstrated that there is no statistically significant difference between automatic and manual transmissions in terms of fuel efficiency (MPGs) and hence the second question regarding quantifying the MPG difference is not able to be answered.

Analysis

Several linear models were fit, however I will only show and describe two since they summarize the trends I observed. The first LM fits MPG against transmission type (factor variable), weight, and engine displacement. The second model fits MPG against transmission type (factor variable), weight, engine displacement, # of forward gears (factor), and # of carborators (factor).

```
fit1 <- lm(mpg ~ factor(am) + wt + disp - 1, data=mtcars)
summary(fit1)
```

```
##
## Call:
## lm(formula = mpg ~ factor(am) + wt + disp - 1, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4890 -2.4106 -0.7232  1.7503  6.3293
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## factor(am)0 34.675911   3.240609  10.700 2.12e-11 ***
## factor(am)1 34.853635   2.376440  14.666 1.14e-14 ***
## wt          -3.279044   1.327509  -2.470   0.0199 *
## disp        -0.017805   0.009375  -1.899   0.0679 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.967 on 28 degrees of freedom
## Multiple R-squared:  0.9824, Adjusted R-squared:  0.9799
## F-statistic: 391.7 on 4 and 28 DF,  p-value: < 2.2e-16
```

```
fit2 <- lm(mpg ~ factor(am) + wt + disp + factor(gear) + factor(carb) -1, data=mtcars)
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ factor(am) + wt + disp + factor(gear) + factor(carb) -
##     1, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0779 -1.2177  0.0671  0.7668  4.8639
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## factor(am)0   28.277817   3.926452   7.202 4.25e-07 ***
## factor(am)1   29.467976   3.604277   8.176 5.80e-08 ***
## wt            -2.036824   1.731270  -1.176   0.2526
## disp          -0.004617   0.016215  -0.285   0.7787
## factor(gear)4  3.633903   3.033237   1.198   0.2443
## factor(gear)5  2.909375   3.281778   0.887   0.3854
## factor(carb)2 -1.597309   1.741439  -0.917   0.3694
## factor(carb)3 -2.842422   2.202439  -1.291   0.2109
## factor(carb)4 -5.225352   2.235957  -2.337   0.0294 *
## factor(carb)6 -6.365944   3.785618  -1.682   0.1075
## factor(carb)8 -8.716297   4.241707  -2.055   0.0525 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.814 on 21 degrees of freedom
## Multiple R-squared:  0.9882, Adjusted R-squared:  0.982
## F-statistic: 159.3 on 11 and 21 DF,  p-value: < 2.2e-16
```

The first model is looking at the important cofounders (weight and engine displacement) and the fit demonstrates that the estimated MPG for an automatic transmission is 34.68 while for a manual transmission it is 34.85 with standard deviations of 3.24 and 2.38 mpgs, respectively. The difference in the means is less than 0.2 mpgs between the two transmissions. Relative to the standard deviations there is not enough gap for there to be statistical significance. This is as expected because the cars with automatic transmissions on average had engines with twice as much displacement and were generally 1300lbs heavier. There was very little overlap of cars with both manual and automatic transmissions that had the similar weights and engine sizes.

Appendix

Data Exploration

Prior to beginning the analysis the first step was to get familiar with the mtcars dataset. I started by loading the dataset, looking at the help for it and then doing a summary of the data. I also decided to look at the dataframe structure with (str) and look at a pairs plot of the variables.

```r
data(mtcars)
?mtcars
```

```
## starting httpd help server ... done
```

```r
summary(mtcars)
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
##        am              gear            carb      
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
##  Median :0.0000   Median :4.000   Median :2.000  
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812  
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

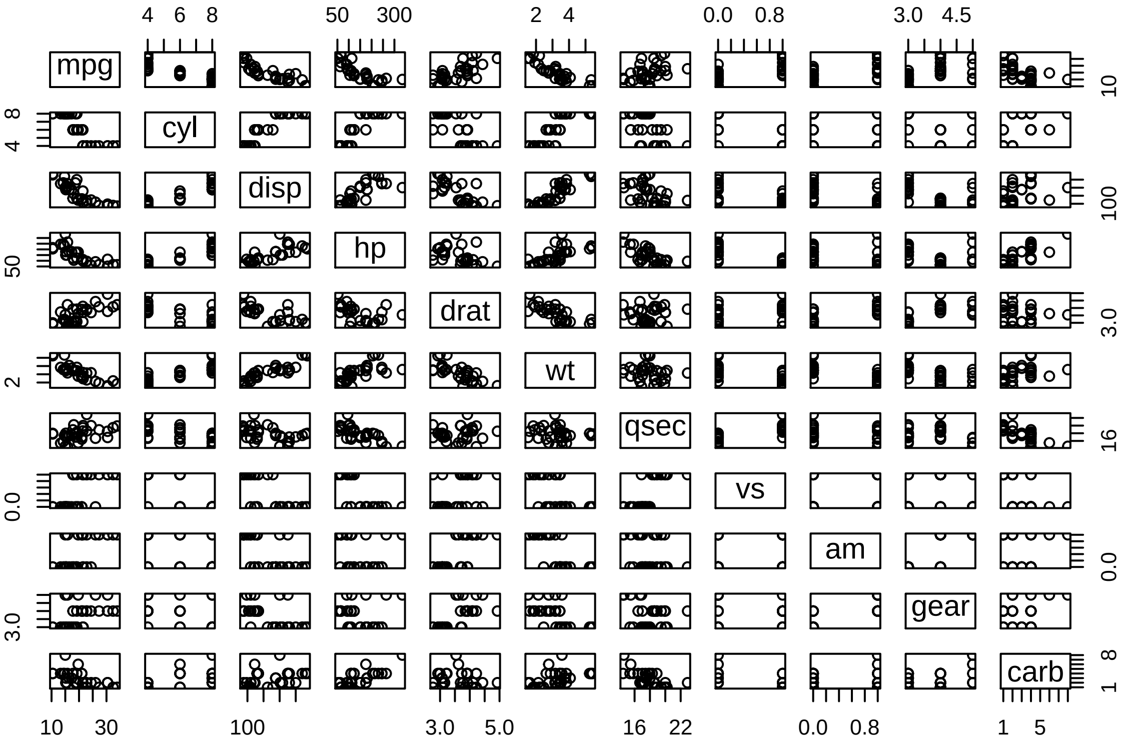
```r
str(mtcars)
```

```
## '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 ...
```

```r
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
```

```r
plot(mtcars)
```



Next, I decided to look at the number of automatic/manual transmissions in the dataset and difference in means between the two types of transmissions

```r
table(mtcars$am)
```

```
## 
##  0  1 
## 19 13
```

```r
tapply(mtcars$mpg, mtcars$am, mean)
```

```
##        0        1 
## 17.14737 24.39231
```

From just looking at the means it seems like manual transmissions are more efficient by a large gap. In order to investigate further I was curious to see if manual cars potentially were heavier or generally had larger engines

```r
tapply(mtcars$disp, mtcars$am, mean)
```

```
##        0        1 
## 290.3789 143.5308
```

```
tapply(mtcars$wt, mtcars$am, mean)
```

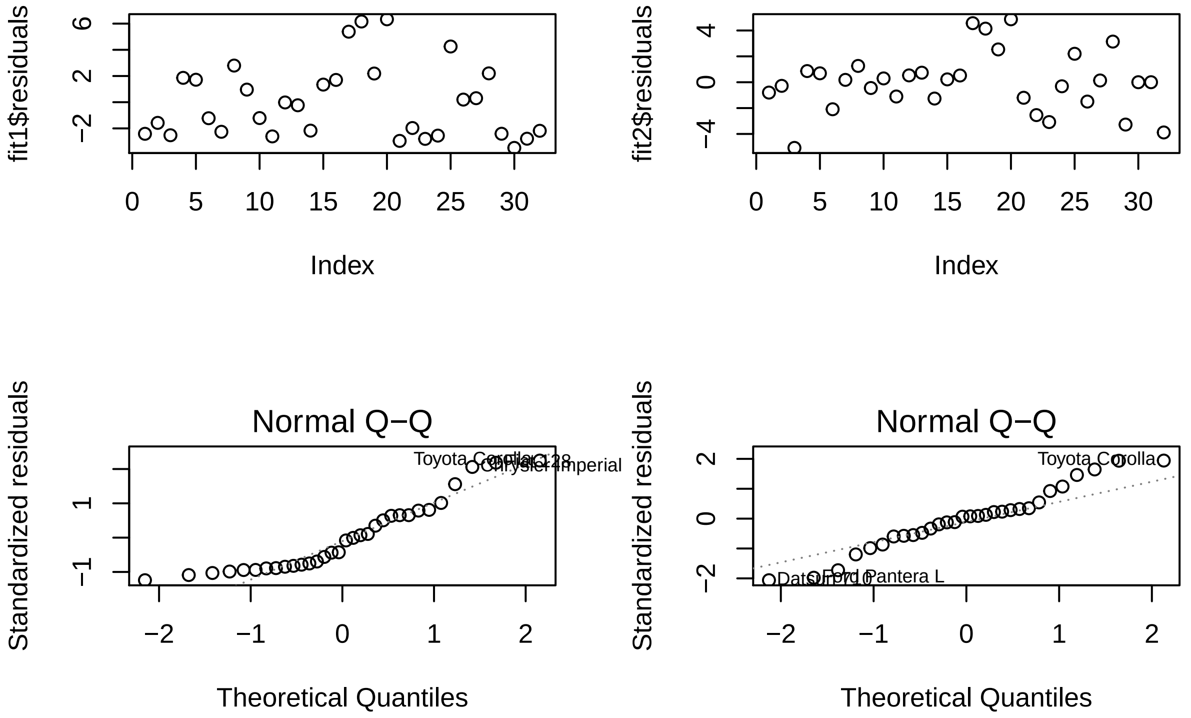
```
##        0        1 
## 3.768895 2.411000
```

Here we see that on average the engine displacement of the automatic transmission cars is roughly double and the car weight on average is 1300lbs heavier so it seems plausible that with heavier cars and bigger engines alone the fuel efficiency (mpgs) will be worse.

Validation

One important thing is doing a residuals plot to understand if the errors appear to meet the general assumption of a normal distribution.

```
## Warning: not plotting observations with leverage one:
##   30, 31
```



From the two plots we can see the residuals. The first fit (based only on transmission, weight and displacement) appears to have several outliers with values of ~6 while the smallest values are under -4. This is an indication it is not a great fit. The second fit (fit2) has a much more traditional scatter of points. In both cases, the residuals are rather and the r-squared value indicates that the fit isn't a great fit