Is An Automatic or Manual Transmission Better For MPG

Richard Ashley

Sunday, September 27, 2015

Executive Summary

The goal of the paper is to quantify the impact of the transmission type [automatic, manual] on the gass mialage [mpg] of the cars that are in the `mtcars` data set. We will be building a multiple regression linear model to identify what factors are improtant to `mpg` so that we can appropriatly comment on impact of transmission type given the impact of other important factors. We found that a manual transmition provided on average a statitically significant benefit of 1.8 mpg more that automatic transmitions although there are other factors that have even more impact on mpg.

Exploratory Data Analyses

We will first load the data set using the `data(mtcars)` command and then eplore what this data set contains.

```
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 ...
```

As we can see, there are 10 factor variables plus `mpg`. A description of the meaning of these variables is given in the appendix. We also perform a `pairs` plot to further look at the relationships between all factors at once. see Figure 1 in the appendix.

From this we can identify that a number of factors appear to have descreat levels and we should code them specifically as a factor for this analysis. This was done with the `mtcars$cyl <- factor(mtcars$cyl)` style command for the facotrs cyl, vs, am, gear, and carb.

It is also clear from Figure 1 that a number of these factors are potentually correlated to `mpg`. For example `disp` and `wt` but that these may also be correlated with each other. We will have to be carefull in our model building to not select to many confounding variables.

We will also quickly look at just a simple box plot of mpg vs. transmission type. This indicates an average of about 7 mpg more for manual transmitions but this does not include any other potential factors. See Figure 2 for the box plot

Modle Selection

To find the best model for determining the impact of transmission type on mpg, we will first look at a model that selects all factors and a model the selects on transmission type. Finally we will then use a stepwise model selection techniche to fit the best model and then look at how much incluence transmision type has with all other significant factors removed.

The simplest model would be to only use `am` as the single explanitory variable. This produces results in Table 1. The basic model does show a significan effect from the transmission variable of 7.2449393. If we then add all factors to the model we get Table 2. Notice that none of the coeficients in this case are significant and that we have over fitted the model. Also from Table 2 you can see the VIF scores and that they are not very unifom. A number of these factors are adding significantly to the variation of the model.

If we then use a stepwise search for the better model using the `setp` function, we get a model as in Table 3. Here the model is mpg ~ cyl + hp + wt + am. In this case the VIF scores are very uniform and the residurl plot below looks resonable with fairly normal residuals.

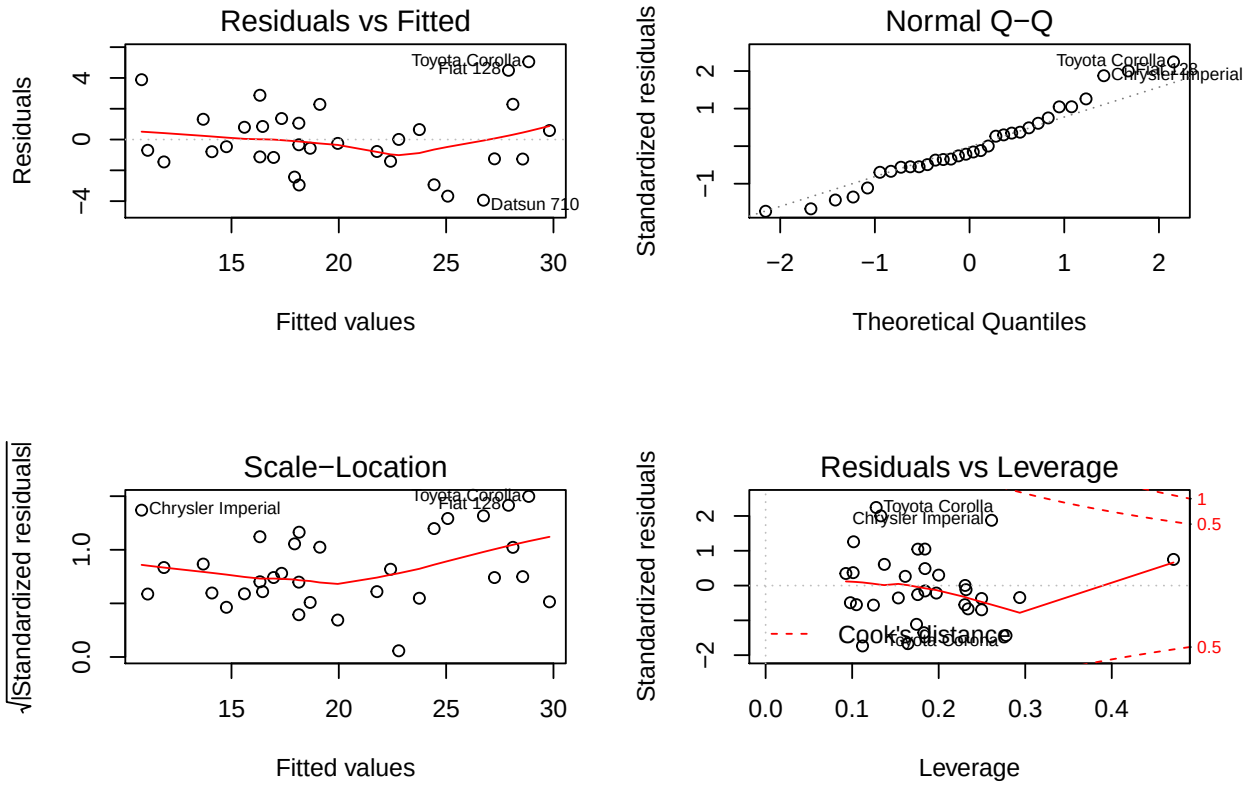

Figure 3:

Statistical Inference

We can build a simple statistical inference of the estimate of the Simple model using a t statistic. For the results see Table 4. For the Simple model of `mpg ~ am` the confidence interval on the estimate (impact of mpg going from Manual to Automatic) is between -11.2801944 and -3.2096842 mpg. So Manual would be significantly more effecient. [*for the code please see the rmarkdown document* at link]

Apendix

Figure 1 - Pairs plot of all factors

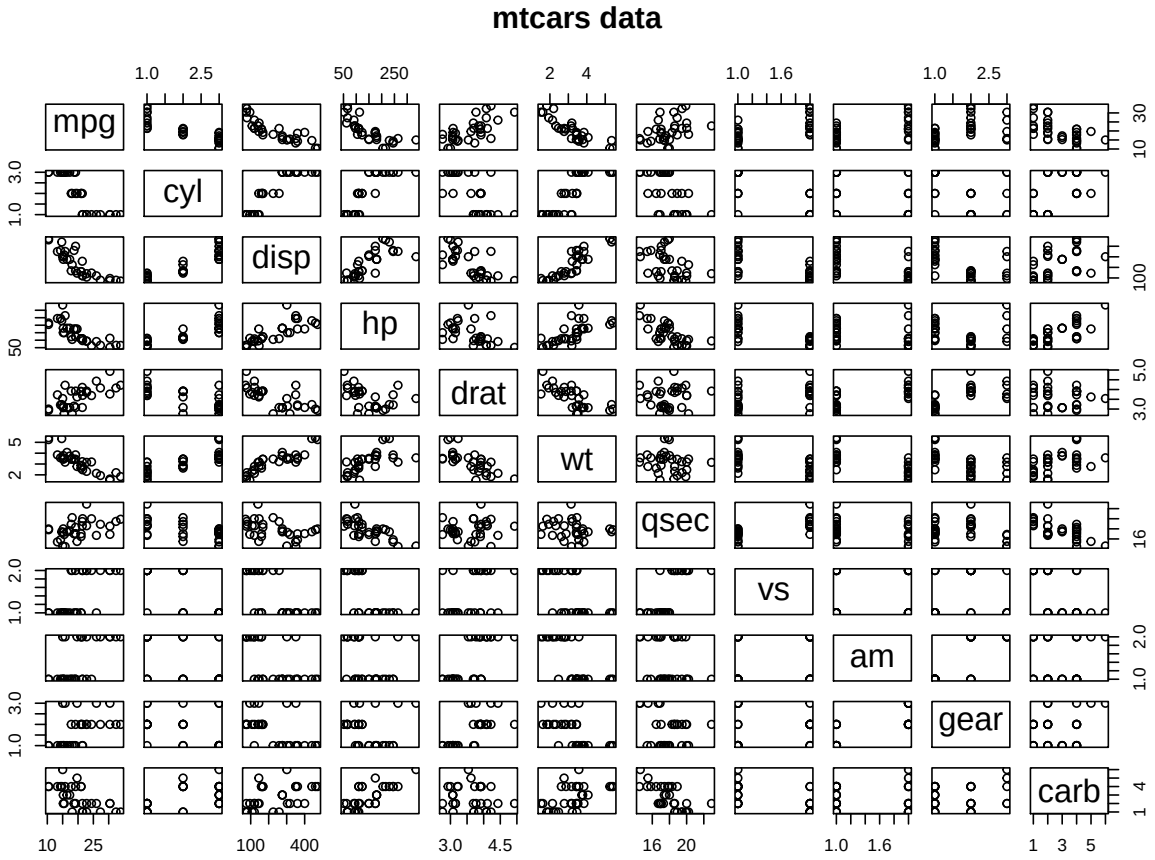

Figure 2 - Boxplot of MPG vs. Transmison Type

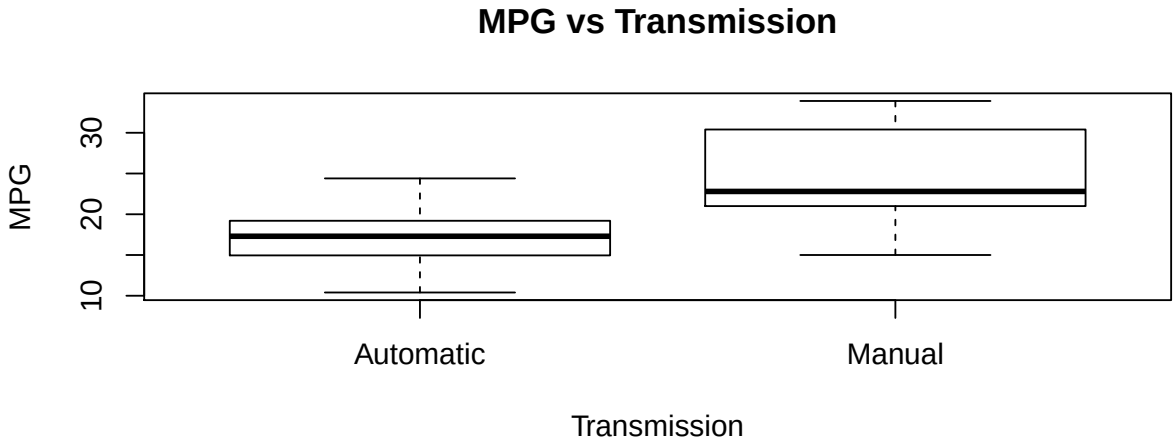

Table 1 - Fit of Simple Model

```
## $coefficients
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## amManual     7.244939   1.764422  4.106127 2.850207e-04
```

Table 2 - Fit of All Model

```
## $coefficients
##                Estimate  Std. Error     t value   Pr(>|t|)
## (Intercept) 23.87913244 20.06582026  1.19004018 0.25252548
## cyl6        -2.64869528  3.04089041 -0.87102622 0.39746642
## cyl8        -0.33616298  7.15953951 -0.04695316 0.96317000
## disp         0.03554632  0.03189920  1.11433290 0.28267339
## hp          -0.07050683  0.03942556 -1.78835344 0.09393155
## drat         1.18283018  2.48348458  0.47627845 0.64073922
## wt          -4.52977584  2.53874584 -1.78425732 0.09461859
## qsec         0.36784482  0.93539569  0.39325050 0.69966720
## vs1          1.93085054  2.87125777  0.67247551 0.51150791
## amManual     1.21211570  3.21354514  0.37718957 0.71131573
## gear4        1.11435494  3.79951726  0.29328856 0.77332027
## gear5        2.52839599  3.73635801  0.67670068 0.50889747
## carb2       -0.97935432  2.31797446 -0.42250436 0.67865093
## carb3        2.99963875  4.29354611  0.69863900 0.49546781
## carb4        1.09142288  4.44961992  0.24528452 0.80956031
## carb6        4.47756921  6.38406242  0.70136677 0.49381268
## carb8        7.25041126  8.36056638  0.86721532 0.39948495
```

```
##            GVIF Df GVIF^(1/(2*Df))
## cyl  128.120962  2        3.364380
## disp  60.365687  1        7.769536
## hp    28.219577  1        5.312210
## drat   6.809663  1        2.609533
## wt    23.830830  1        4.881683
## qsec  10.790189  1        3.284842
## vs     8.088166  1        2.843970
## am     9.930495  1        3.151269
## gear  50.852311  2        2.670408
## carb 503.211851  5        1.862838
```

Table 3 - Fit of AIC Stepwise Model, mpg ~ cyl + hp + wt + am

```
## $coefficients
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 33.70832390 2.60488618 12.940421 7.733392e-13
## cyl6        -3.03134449 1.40728351 -2.154040 4.068272e-02
## cyl8        -2.16367532 2.28425172 -0.947214 3.522509e-01
## hp          -0.03210943 0.01369257 -2.345025 2.693461e-02
## wt          -2.49682942 0.88558779 -2.819404 9.081408e-03
## amManual     1.80921138 1.39630450  1.295714 2.064597e-01
```

```
##         GVIF Df GVIF^(1/(2*Df))
## cyl 5.824545  2        1.553515
## hp  4.703625  1        2.168784
## wt  4.007113  1        2.001778
## am  2.590777  1        1.609589
```

Table 4 - Statistical Inference, Confiedence intermal of Basic model

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## 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 in group Automatic    mean in group Manual
##                17.14737                24.39231
```

Variable Discriptions

variable	Discription
mpg	Miles/(US) gallon
cyl	Number of cylinders
disp	Displacement (cu.in.)
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (lb/1000)
qsec	1/4 mile time
vs	V/S
am	Transmission (0 = automatic, 1 = manual)
gear	Number of forward gears
carb	Number of carburetors