Is MPG gain better with auto or manual transmission ?

Chidambaranathan, Kavitha

19-05-2017

Executive Summary:

Provided with a data set of 32 different car types in mtcars, the aim was to find whether automatic transmission or manual transmission leads to better gain of miles per gallon for a car. Other significant contributing variables should also be explored.

It was found that manual transmission clearly has a better gain to the mpg value than auto transmission. The gain is quantified to about 1.8 mpg using manual transmission.

Exploratory Analysis:

Data definition provided for the mtcars data set is as follows: A data frame with 32 observations on 11 variables.
1. mpg Miles/(US) gallon
2. cyl Number of cylinders
3. disp Displacement (cu.in.)
4. hp Gross horsepower
5. drat Rear axle ratio
6. wt Weight (1000 lbs)
7. qsec 1/4 mile time
8. vs V/S
9. am Transmission (0 = automatic, 1 = manual)
10. gear Number of forward gears
11. carb Number of carburetors

As a first step, check sample data of mtcars and their data type.

```
library(car)
data(mtcars)
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
```

```
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 ...
```

Some of the fields will need to change to factor.

```
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$am <- as.factor(mtcars$am)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$gear <- as.factor(mtcars$gear)
mtcars$carb <- as.factor(mtcars$carb)
```

Model Selection

In order to determine the variables to include in the model, we first fit the linear model by including all variables and checking the vif.

```
fitall <- lm(mpg ~ ., data = mtcars)
vif(fitall)
```

```
##            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
```

Including many variables that are highly correlated to *am*, our variable of interest should be avoided. Top 3 variables with high correlation here are carb, cyl and disp and need to be watched out for. I would next use the step function to work backwards for model selection.

```
step (fitall, direction = "backward")
```

```
## Start:  AIC=76.4
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - carb  5   13.5989 134.00 69.828
## - gear  2    3.9729 124.38 73.442
## - am    1    1.1420 121.55 74.705
## - qsec  1    1.2413 121.64 74.732
## - drat  1    1.8208 122.22 74.884
## - cyl   2   10.9314 131.33 75.184
## - vs    1    3.6299 124.03 75.354
## <none>              120.40 76.403
## - disp  1    9.9672 130.37 76.948
## - wt    1   25.5541 145.96 80.562
## - hp    1   25.6715 146.07 80.588
##
## Step:  AIC=69.83
```

```
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear
##
##        Df Sum of Sq    RSS    AIC
## - gear  2    5.0215 139.02 67.005
## - disp  1    0.9934 135.00 68.064
## - drat  1    1.1854 135.19 68.110
## - vs    1    3.6763 137.68 68.694
## - cyl   2   12.5642 146.57 68.696
## - qsec  1    5.2634 139.26 69.061
## <none>              134.00 69.828
## - am    1   11.9255 145.93 70.556
## - wt    1   19.7963 153.80 72.237
## - hp    1   22.7935 156.79 72.855
##
## Step:  AIC=67
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - drat  1    0.9672 139.99 65.227
## - cyl   2   10.4247 149.45 65.319
## - disp  1    1.5483 140.57 65.359
## - vs    1    2.1829 141.21 65.503
## - qsec  1    3.6324 142.66 65.830
## <none>              139.02 67.005
## - am    1   16.5665 155.59 68.608
## - hp    1   18.1768 157.20 68.937
## - wt    1   31.1896 170.21 71.482
##
## Step:  AIC=65.23
## mpg ~ cyl + disp + hp + wt + qsec + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - disp  1    1.2474 141.24 63.511
## - vs    1    2.3403 142.33 63.757
## - cyl   2   12.3267 152.32 63.927
## - qsec  1    3.1000 143.09 63.928
## <none>              139.99 65.227
## - hp    1   17.7382 157.73 67.044
## - am    1   19.4660 159.46 67.393
## - wt    1   30.7151 170.71 69.574
##
## Step:  AIC=63.51
## mpg ~ cyl + hp + wt + qsec + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - qsec  1     2.442 143.68 62.059
## - vs    1     2.744 143.98 62.126
## - cyl   2    18.580 159.82 63.466
## <none>              141.24 63.511
## - hp    1    18.184 159.42 65.386
## - am    1    18.885 160.12 65.527
## - wt    1    39.645 180.88 69.428
##
## Step:  AIC=62.06
```

```
## mpg ~ cyl + hp + wt + vs + am
##
##         Df Sum of Sq    RSS    AIC
## - vs    1     7.346 151.03 61.655
## <none>              143.68 62.059
## - cyl   2    25.284 168.96 63.246
## - am    1    16.443 160.12 63.527
## - hp    1    36.344 180.02 67.275
## - wt    1    41.088 184.77 68.108
##
## Step:  AIC=61.65
## mpg ~ cyl + hp + wt + am
##
##         Df Sum of Sq    RSS    AIC
## <none>              151.03 61.655
## - am    1     9.752 160.78 61.657
## - cyl   2    29.265 180.29 63.323
## - hp    1    31.943 182.97 65.794
## - wt    1    46.173 197.20 68.191
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Coefficients:
## (Intercept)         cyl6         cyl8           hp           wt
##    33.70832     -3.03134     -2.16368     -0.03211     -2.49683
##         am1
##     1.80921
```

Step analyses 7 models and returns the best model fit as cyl + hp + wt + am

As the differences in AIC for the last 3 models are less than 1 in each case,it bears to check these further with anova to see the RSS.

```r
fit3 = lm(mpg ~ am + cyl + hp + wt + qsec + vs, data=mtcars)
fit2 = lm(mpg ~ am + cyl + hp + wt + vs, data=mtcars)
fit1 = lm(mpg ~ am + cyl + hp + wt, data=mtcars)
```

```r
anova(fit3, fit2, fit1)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am + cyl + hp + wt + qsec + vs
## Model 2: mpg ~ am + cyl + hp + wt + vs
## Model 3: mpg ~ am + cyl + hp + wt
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     24 141.24
## 2     25 143.68 -1   -2.4420 0.4150 0.5256
## 3     26 151.03 -1   -7.3459 1.2483 0.2749
```

Model3 still looks to be be best fit. Lets check the co-efficients of this model.

```r
bestfit <- lm(formula = mpg ~ am + cyl + hp + wt, data = mtcars)
summary(bestfit)$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 33.70832390 2.60488618 12.940421 7.733392e-13
```

```
## am1          1.80921138 1.39630450  1.295714 2.064597e-01
## 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
```

I would tend to go with the bestfit model to analyze the graphs further. See appendix for the graphs.

The residual vs. fitted plot looks as expected with most of the values lying above and below the 0 reference line. No major pattern such as heteroskedacity is observed. The QQ plot also shows that data clearly fall on the regression line proving normality of errors.

Selected Model Analysis

```r
vif(bestfit)
```

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

```r
summary(bestfit)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + hp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## am1          1.80921    1.39630   1.296  0.20646
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

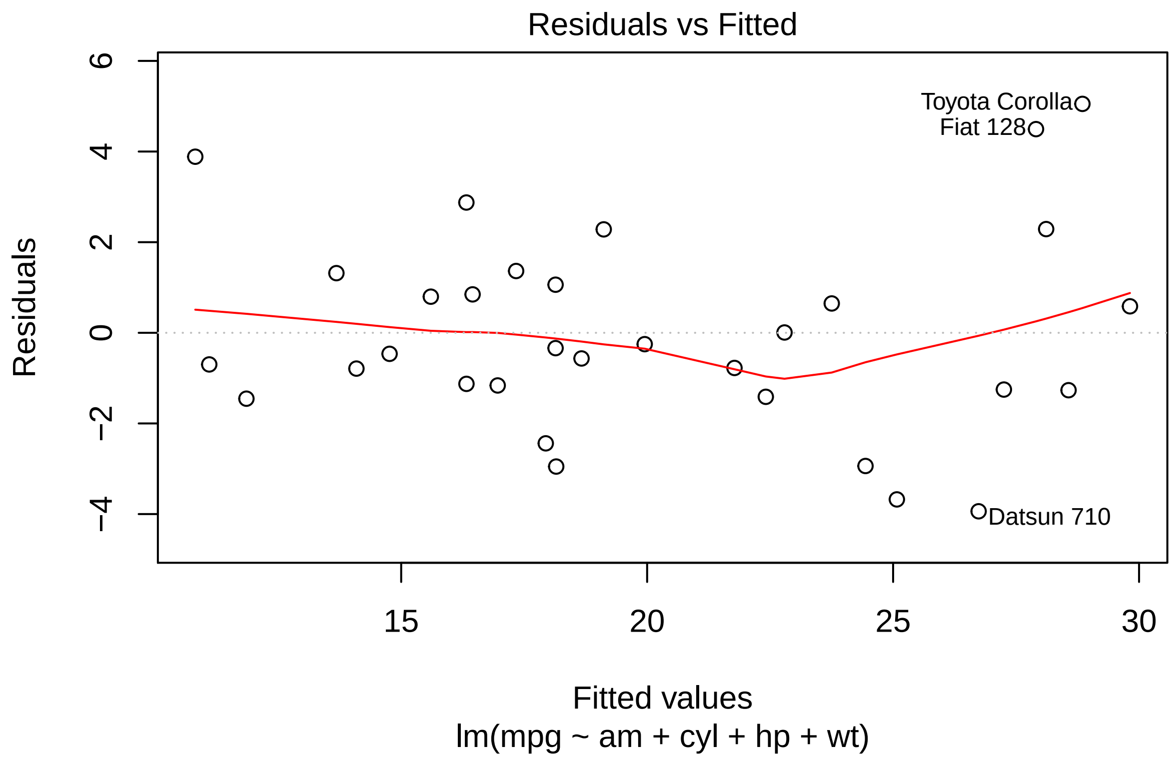
Conclusion

MPG gain for manual transmission is 1.80. The probability is within acceptable range of 0.2. It should be noted that manual transmission is not the only factor contributing to mpg. There could be other factors including those not observed. However, it can be said number of cylinders, horse power and weight of the car also significantly contribute to the mpg value.

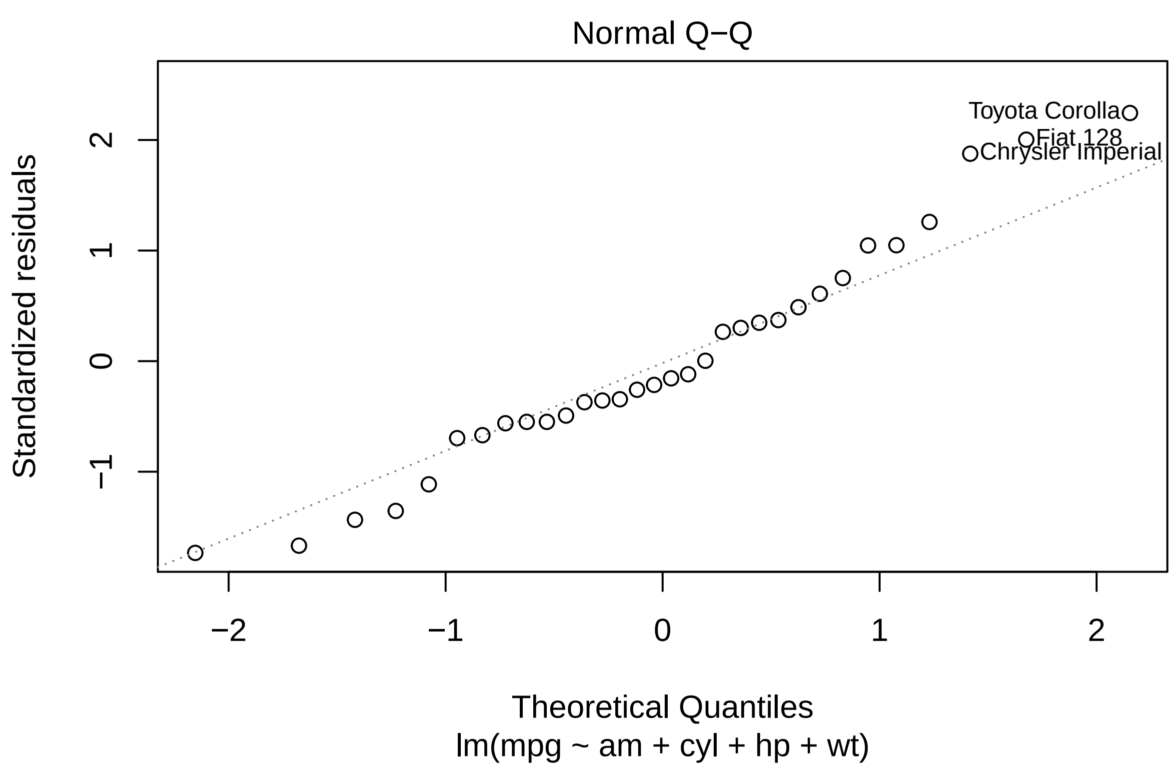
Appendix

Graphs ploted for the best fit model.

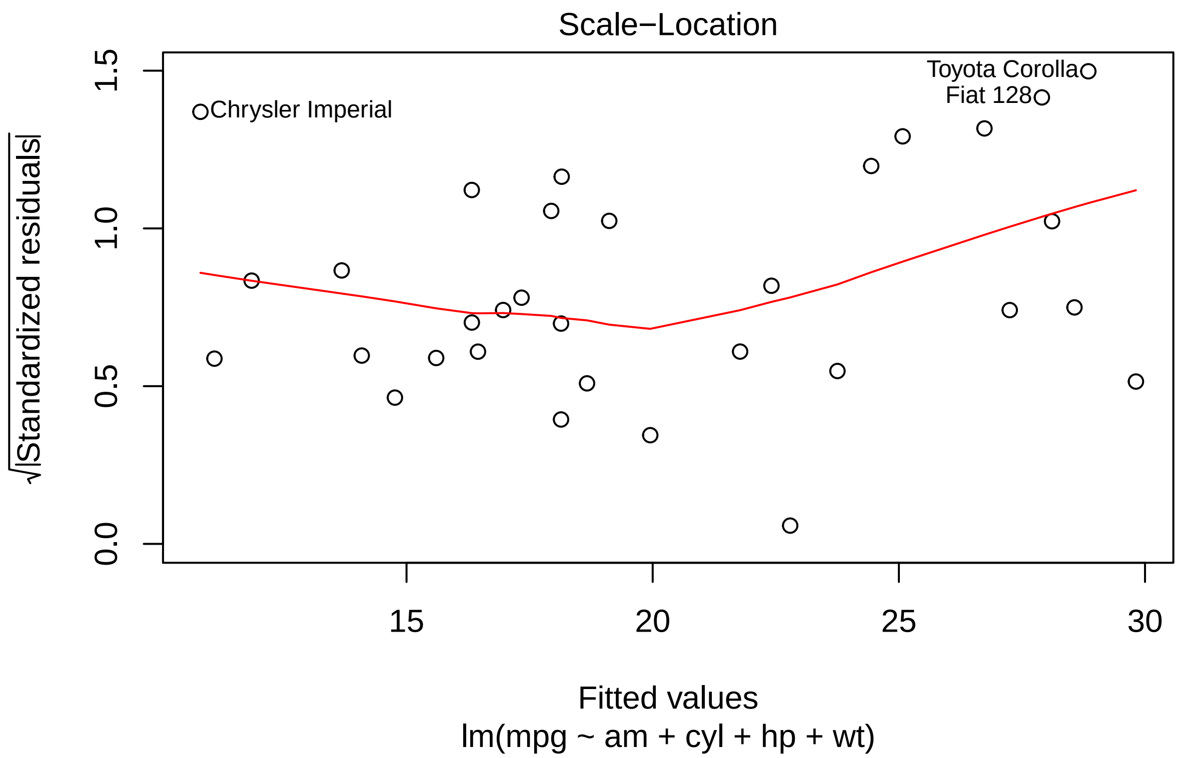
```r
plot(bestfit)
```

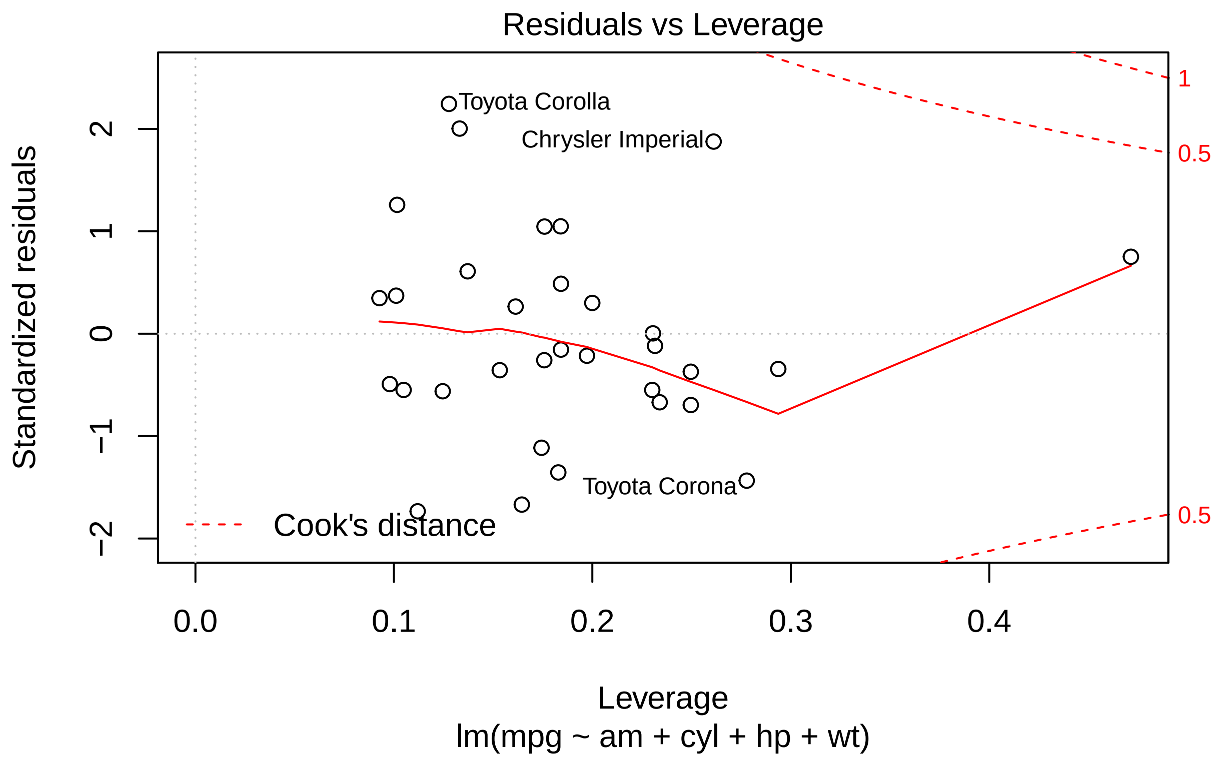


Residuals vs Fitted
Toyota Corolla
Fiat 128
Datsun 710
Residuals
Fitted values
lm(mpg ~ am + cyl + hp + wt)



Normal Q−Q
Toyota Corolla
Fiat 128
Chrysler Imperial
Standardized residuals
Theoretical Quantiles
lm(mpg ~ am + cyl + hp + wt)


Scale−Location
Toyota Corolla
Fiat 128
Chrysler Imperial
√|Standardized residuals|
0.0
0.5
1.0
1.5
15
20
25
30
Fitted values
lm(mpg ~ am + cyl + hp + wt)


Residuals vs Leverage
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
1
0.5
0.5
Standardized residuals
2
1
0
−1
−2
0.0
0.1
0.2
0.3
0.4
Leverage
lm(mpg ~ am + cyl + hp + wt)