Relation between miles per gallon and gear transmission type

Harish Kumar Rongala

January 20, 2017

1. Executive Summary

This document explores the relationship between **miles per gallon** (MPG) and factors affecting it. We used "mtcars" data set from *datasets* package in R. It is extracted from the 1974 Motor Trend US magazine, which is about the automobile industry. It comprises fuel consumption and **10** other aspects of automobile design and performance for **32** automobiles (1973-74 models).

This document attempts to answer the following questions

- Which provides better MPG, Automatic or manual transmission ?
- Quantify the MPG difference between automatic and manual transmissions

2. Exploratory data analysis

After looking at individual variable's relationship with **mpg** [see appendix-6.1], we can find significant trends in the plots, showing a positive or negative impact on the mpg. In the **regression modeling section** we will quantify this relationship.

The following 5 variables are factor variables but they are labeled as a numeric class. We have to transform these variables in to factor class to make more sense of them in our modeling.

```
## [1] "cyl"  "vs"   "am"   "gear" "carb"
```

```
mtcars$cyl<-factor(mtcars$cyl);
mtcars$vs<-factor(mtcars$vs);
mtcars$am<-factor(mtcars$am);
levels(mtcars$am)<-c("automatic","manual");
mtcars$gear<-factor(mtcars$gear);
mtcars$carb<-factor(mtcars$carb);
```

3. Regression modeling

As the dependent variable **mpg** is not binomial or a count variable, we use a linear model to fit our data. In our first model we include all the independent variables in the model, as we found that they have certain degree of impact on mpg from exploratory data analysis.

```
## Output of this fit can be found in appendix #Fit0
fit0<-lm(mpg~.,data=mtcars);
round(summary(fit0)$coef,3);
```

P-values of most variables are insignificant, so we drop those variables and re-fit the model. However, we can't drop **am** variable which indicates transmission mode. Because we want to find the relation between **am** and **mpg**. In our second model, our independent variables are **am, hp and wt**.

```
## Output of this fit can be found in appendix #Fit1
fit1<-lm(mpg~am+hp+wt,data=mtcars);
round(summary(fit1)$coef,3);
```

We use a step wise model selection algorithm based on **AIC** Akaike's 'An Information Criterion' to fit a new linear model.

```
better_fit<-step(fit0,direction = "both");
```

Now, we got 3 models. However, we knew that first model - fit0 is too insignificant when compared to other two models. We use **ANOVA** variance analysis technique to analyze our models.

```
anova(fit1,better_fit);
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am + hp + wt
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     28 180.29
## 2     26 151.03  2    29.265 2.5191    0.1 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our anova test shows that model 3 - 'better_fit' is significant than 'fit1', with a P-value of **0.1**. Let's look at the coefficients given by this model.

```
round(summary(better_fit)$coef,5);
```

```
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 33.70832    2.60489 12.94042  0.00000
## cyl6        -3.03134    1.40728 -2.15404  0.04068
## cyl8        -2.16368    2.28425 -0.94721  0.35225
## hp          -0.03211    0.01369 -2.34503  0.02693
## wt          -2.49683    0.88559 -2.81940  0.00908
## ammanual     1.80921    1.39630  1.29571  0.20646
```

4. Inference & Conclusion

- Cars with **Manual** transmission will have more miles per (US) gallon by a factor of **1.809**, over an **automatic** transmission. This value is adjusted by considering horsepower, weight and no. of cylinders.
- **95%** of this factor lies between **-1.06 to 4.67** (see Appendix 6.3.4), as zero lies in the interval, we are not quite confident about its significance.
- Our linear model explains **86.58%** of the total variance.

5. Diagnostics

Refer appendix 6.2, to find the diagnostics plot of the 'better_fit' model.

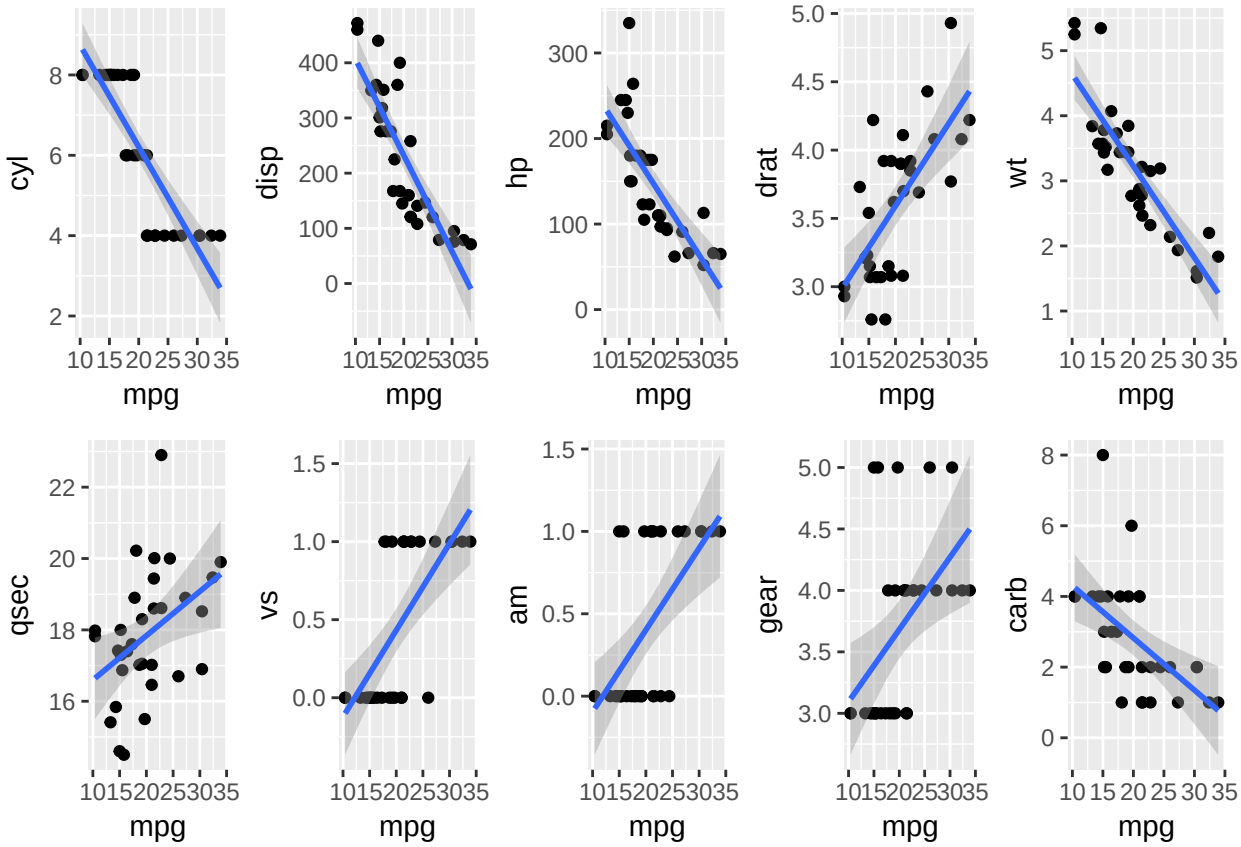
- Residual Vs Fitted plot doesn't show any **systematic pattern** or **heteroskedasticity**, so our model fit is good.
- Remaining plots show few outlier which have some **influence and leverage** on the model fit. The following car models are the outliers

```
df<-as.data.frame(dfbetas(better_fit));
rownames(df[df$ammanual %in% tail(sort(df$ammanual),4),]);
```

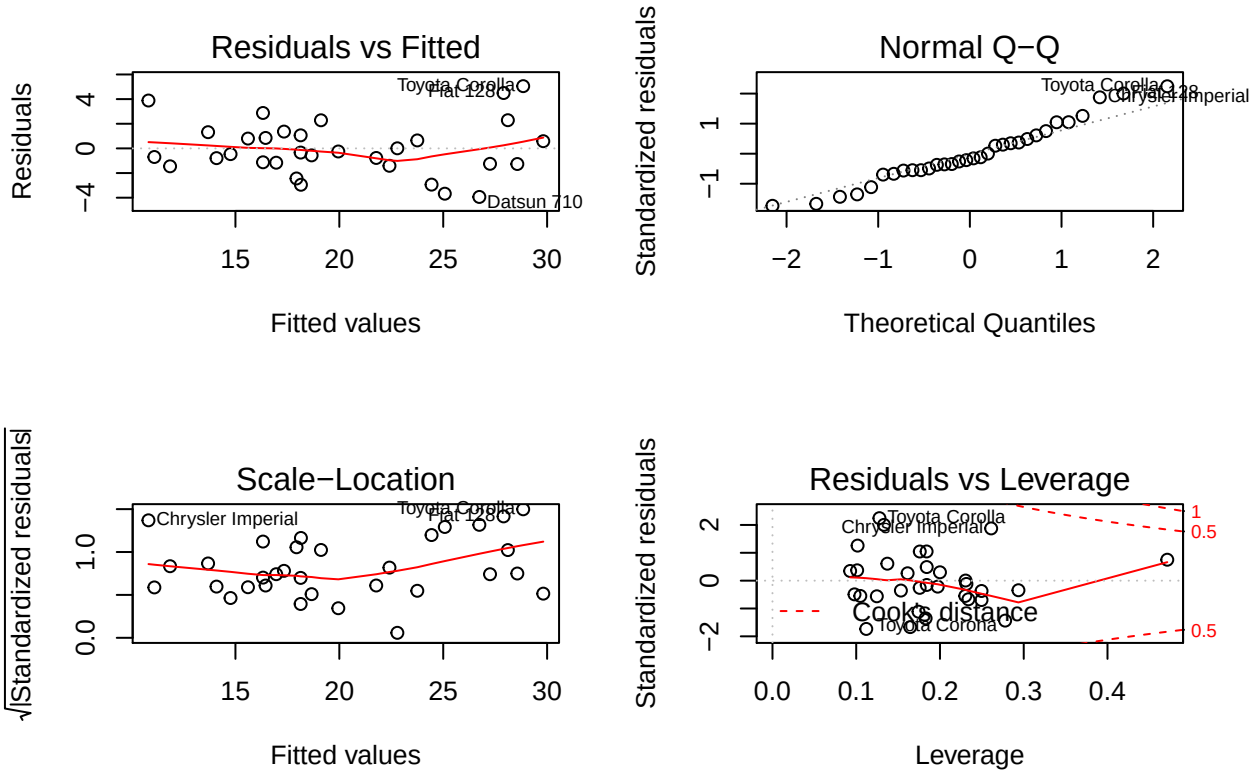
```
## [1] "Chrysler Imperial" "Fiat 128"          "Toyota Corolla"
## [4] "Toyota Corona"
```

6. Appendix

6.1. Relationship plots



6.2. Diagnostics plot



6.3. Summary of linear models

6.3.1. First model - fit0

```
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 23.87913   20.06582  1.19004  0.25253
## cyl6        -2.64870    3.04089 -0.87103  0.39747
## cyl8        -0.33616    7.15954 -0.04695  0.96317
## disp         0.03555    0.03190  1.11433  0.28267
## hp          -0.07051    0.03943 -1.78835  0.09393
## drat         1.18283    2.48348  0.47628  0.64074
## wt          -4.52978    2.53875 -1.78426  0.09462
## qsec         0.36784    0.93540  0.39325  0.69967
## vs1          1.93085    2.87126  0.67248  0.51151
## ammanual     1.21212    3.21355  0.37719  0.71132
## gear4        1.11435    3.79952  0.29329  0.77332
## gear5        2.52840    3.73636  0.67670  0.50890
## carb2       -0.97935    2.31797 -0.42250  0.67865
## carb3        2.99964    4.29355  0.69864  0.49547
## carb4        1.09142    4.44962  0.24528  0.80956
## carb6        4.47757    6.38406  0.70137  0.49381
## carb8        7.25041    8.36057  0.86722  0.39948
```

6.3.2. Second model - fit1

```
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 34.00288    2.64266 12.86692  0.00000
## ammanual     2.08371    1.37642  1.51386  0.14127
## hp          -0.03748    0.00961 -3.90183  0.00055
## wt          -2.87858    0.90497 -3.18085  0.00357
```

6.3.3. Third model - better_fit

```
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 33.70832    2.60489 12.94042  0.00000
## cyl6        -3.03134    1.40728 -2.15404  0.04068
## cyl8        -2.16368    2.28425 -0.94721  0.35225
## hp          -0.03211    0.01369 -2.34503  0.02693
## wt          -2.49683    0.88559 -2.81940  0.00908
## ammanual     1.80921    1.39630  1.29571  0.20646
```

6.3.4. 95% Confidence interval

```
# Get the 95% confidence interval
confint(better_fit,'ammanual');
```

```
##              2.5 %   97.5 %
## ammanual -1.060934 4.679356
```