Relation between miles per gallon and gear transmission type

Harish Kumar Rongala

January 20, 2017

Executive Summary

This document explores the relationship between **miles per gallon** (MPG) and factors that affect it. We used "mtcars" data set. It was extracted from the 1974 Motor Trend US magazine, which is about the automobile industry. It comprises fuel consumption and **10** 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

Exploratory data analysis

After looking at individual variable's relationship with **mpg** [appendix #Relationship plots], there seems to be a trend in the plots, showing positive or negative impact on the mpg. In the **regression models section** we will quantify this relationship.

The following 5 variables are factor variables but they are labeled as 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);
```

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 don't drop **am** variable which indicates transmission mode. Because, we want to find the relation between **am** and **mpg**.

```
## 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'.

```
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,3);
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   33.708      2.605  12.940    0.000
## cyl6          -3.031      1.407  -2.154    0.041
## cyl8          -2.164      2.284  -0.947    0.352
## hp            -0.032      0.014  -2.345    0.027
## wt            -2.497      0.886  -2.819    0.009
## ammanual       1.809      1.396   1.296    0.206
```

Inference & Conclusion

- Cars with **Manual** transmission will have more miles per (US) gallon by a factor of **1.809**, over **automatic** transmission. This value is adjusted by considering horse power, weight and no. of cylinders.
- **95%** of this factor lies between **-1.06 to 4.67**, as zero lies in the interval, we are not quite confident about its significance.

Diagnostics

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

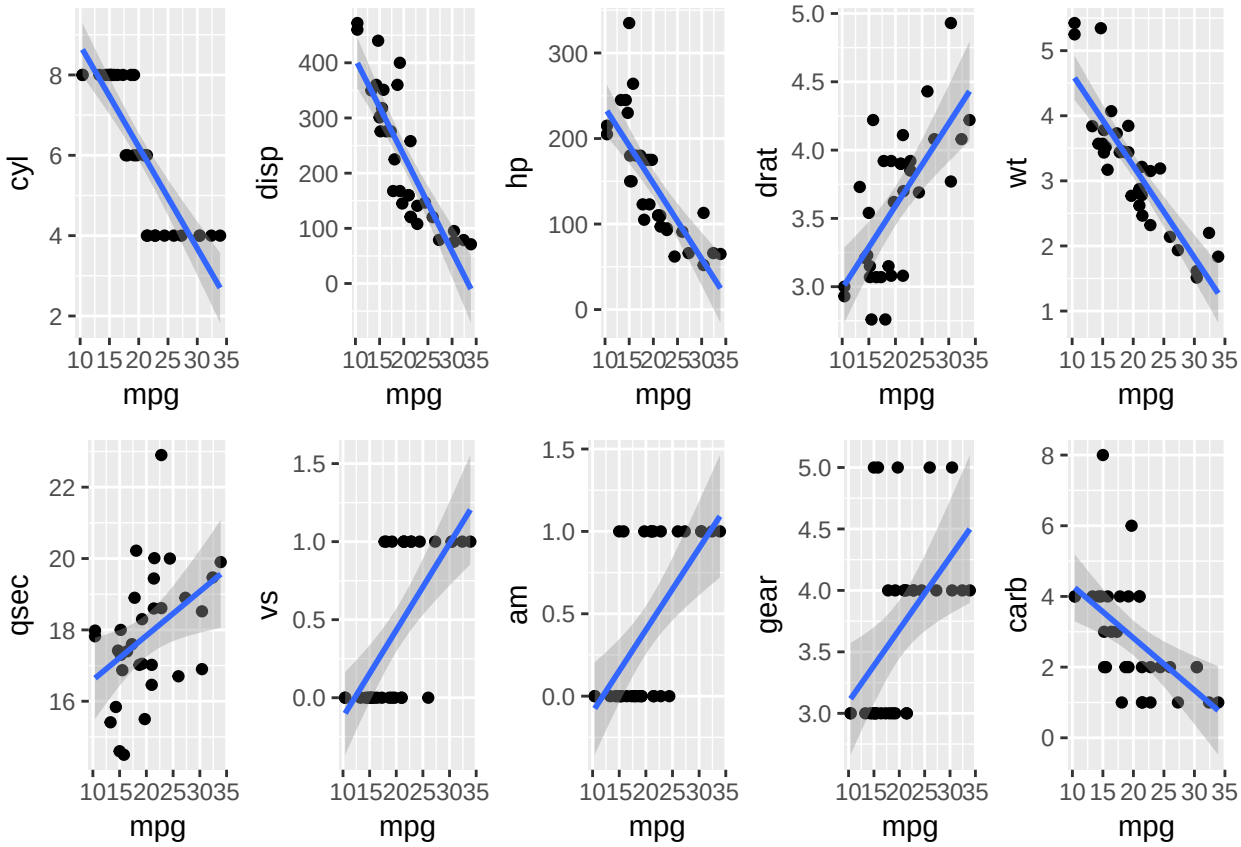
- Residual Vs Fitted plot doesn't show any **systematic pattern** or **heteroskedasticity**, so our model fit is good.
- Remaining plots show few outlier which have some **influence and leverage** on the model fit. The following car models are the outliers

```
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"
```

Appendix

Relationship plots



Diagnostics plot