A Study on Vehicles' Gas Mileage via Regression Models

Wenhe 'Wayne' Ye

September 25, 2015

Note: In order to save space for the report, we put all the plots in the supporting information section.

Executive Summary

We analyzed the data set mtcars with 36 observations. Our approaches to reach the research target are listed below:

1.Are there any confounding variables other than the factor (types of transmission)?
- An exploratory analysis is carried out to determine the confounding terms.

2.How to find a model to describe the correlation between mpg and other parameters?
- Use VIF to detect high correlation variables
- Use ANOVA to judge the models validity

3.How to interpret the fitted coeficients?
- A linear regression model with interaction term is suggested. mpg may also depend on different car weight regime.

Data Processing

First we need to call a few useful R packages to facilitate our analysis and load data *mtcars* into work space.

```r
library(ggplot2);library(dplyr);library(grid);library(gridExtra);library(car);library(xtable)
data(mtcars)
```

Clean up the raw data and convert some varibles into factors.

```r
mtcars2<-mutate(mtcars,mpg,disp,wt,hp,Cylinder=as.factor(cyl),AutoTransmission=as.factor(am))
mtcars2<-select(mtcars2,mpg,disp,wt,hp,Cylinder,AutoTransmission)
```

Exploratory Data Analysis

Since we want to explore the relationship between the mpg and whether the cars are manual or auto transmission. We make a violin plot (in supporting information (SI)) between mpg and factor of different transmission types to see the over all relationship.
From the plot we see manual transmission cars have a higher gas mileage over the automatic transmission. However, there might be other confounding variables need to be taken into account. For example, more high engine displacement cars tend to have manual transmission rather automatics while most economic cars are with manual ones. We select a few other variables as candidates to see their correlation with the mpg data. We picked displacement (disp), horsepower (hp), weight(wt) and number of cylinders(Cylinder) as confounding variables. (Figures can be found in SI), which supports that these variables also show some suspicious correlation with mpg. It is worth noting here, we transform the cylinder number into factors rather than a continuous varible in the following study. In order to quantify the difference between an auto transmission car and a manual transmission car, we need to carefully select the model with suitable variables and factors to make our estimation.

Regression Modeling

Our first attempt is to build a regression model includes all mentioned variables and factors (with no interaction). (mpg ~ AutoTransmission, Cylinder, disp, hp, wt).

```
fit_all<-lm(data=mtcars2,mpg~.)
```

However, the variance inflation factors (VIF) for the *fit_all* model is not optimistic:

```
vif_table<-vif(fit_all)
vif_table[,1]
```

```
##             disp               wt               hp         Cylinder 
##        12.901490         6.821979         4.736101         9.765272 
## AutoTransmission 
##         2.590898
```

Some VIFs have relatively high values indicating some strong correlation between variables/factors. After a trial and error process (we use ANOVA as a tool to judge if the model is under- or overfit). In the end, we only keep factors AutoTransmission and varible wt in the regression model. The ANOVA table below indicates the variables included are sufficient compare to *fit_all* (P-value>0.05, which means we could not reject the null hypothesis).

```
## Analysis of Variance Table
## 
## Model 1: mpg ~ AutoTransmission * wt
## Model 2: mpg ~ AutoTransmission * wt + Cylinder + disp + hp
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)  
## 1     28 188.01                              
## 2     24 130.38  4    57.631 2.6522 0.05786 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Below is the summary of the model we fit.

```
summary(fit_1)
```

```
## 
## Call:
## lm(formula = mpg ~ AutoTransmission * wt, data = mtcars2)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.6004 -1.5446 -0.5325  0.9012  6.0909 
## 
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          31.4161     3.0201  10.402 4.00e-11 ***
## AutoTransmission1    14.8784     4.2640   3.489  0.00162 ** 
## wt                   -3.7859     0.7856  -4.819 4.55e-05 ***
## AutoTransmission1:wt  -5.2984     1.4447  -3.667  0.00102 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.591 on 28 degrees of freedom
## Multiple R-squared:  0.833,  Adjusted R-squared:  0.8151
## F-statistic: 46.57 on 3 and 28 DF,  p-value: 5.209e-11
```

All t-tests suggest signifant correlation in the varibles and factors we choosed. The residual & diagnostics plot to our fitted model also suggests no obvious pattern existed in residual (in SI). The plot with chosen variable and factor is provided in the SI.

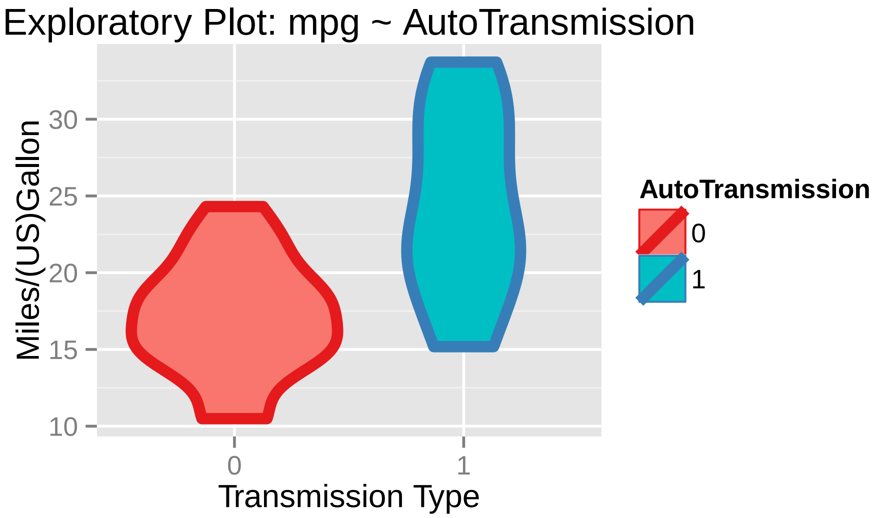
Results

Q1 "Is an automatic or manual transmission better for MPG"
From the data set, we see the mean mpg for manual transmission cars are higher than the automatic ones. However, this comparison should be carried out in a more controlled circumstance. For example, if we are considering the cars within a certain range of weight. A car weighs less than 3000 lbs with automatic tranmission tends to have a higher mpg than a manual transmission car in the same weight region. However, a heavier car (>3000 lbs) with manual transmission is more likely to have a higher mpg than an automatic transmission car.

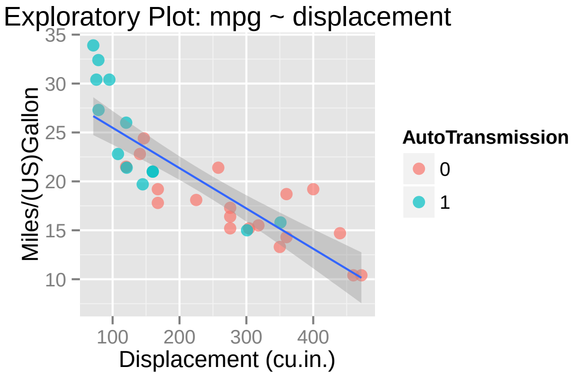
Q2 "Quantify the MPG difference between automatic and manual transmissions" From our fitted model *fit_1*. Given the weight wt in lbs. The gap between a manual transmission car and an automatic transmission car (auto-manual) equals 14.88 - 3.78*wt. which means if a cars weighs more than 14.88/5.30 = 2800 lbs, you might need to choose a manual transmission one to achieve a higher mgp value.

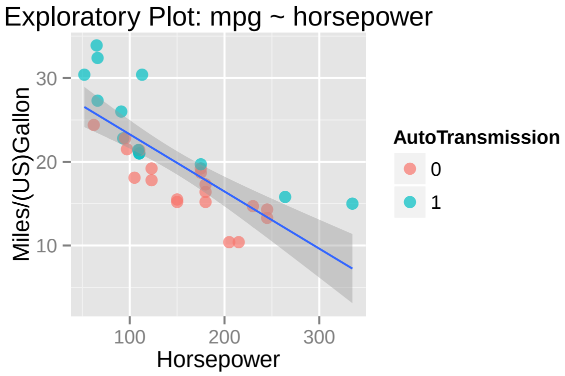
Supporting Information

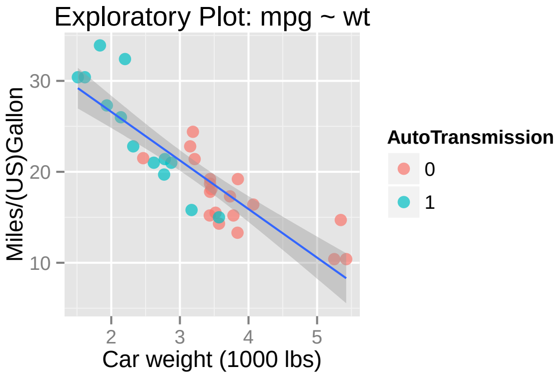
1.mpg vs Transmission Type Violin Plot

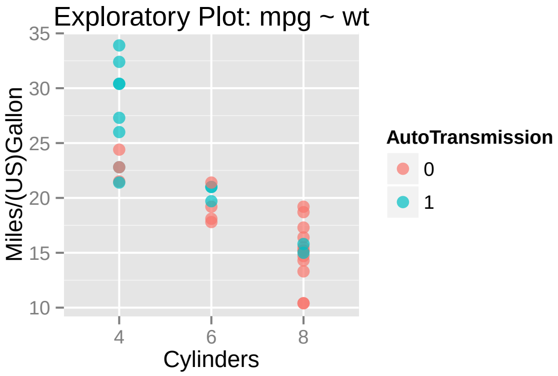


2.mpg vs disp, hp, wt, Cylinder

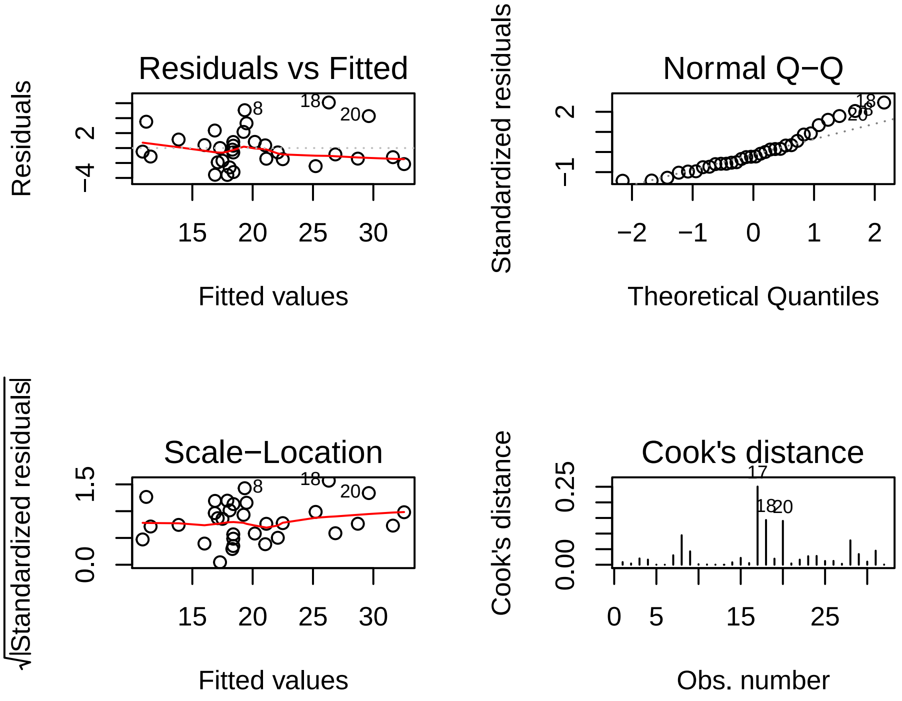






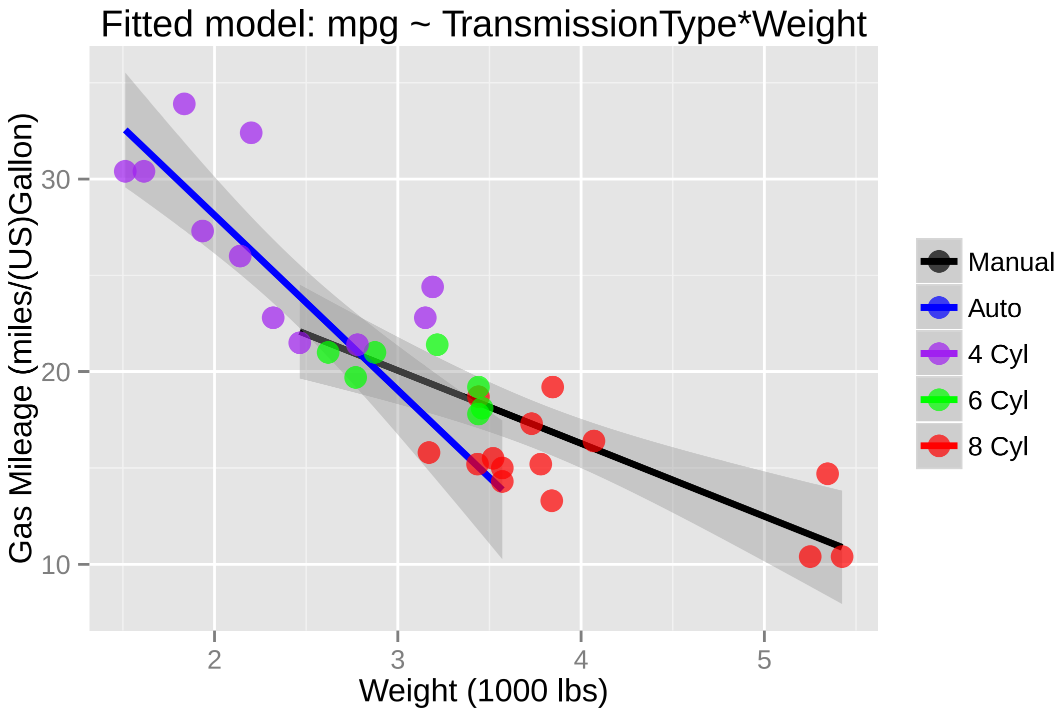


3.Residuals and Diagnostics



4.Fitted model:mpg~Transmission Type * Weight

```
g6<-ggplot(data=mtcars2,aes(x=wt,y=mpg))+
        geom_smooth(method="lm",aes(group=AutoTransmission,col=AutoTransmission),size=1.2)+
        geom_point(size=4,alpha=0.7,aes(col=Cylinder))+
        scale_colour_manual("",values=c("black","blue","purple","green","red"),labels=c("Manual","Auto"
        labs(x="Weight (1000 lbs)",y="Gas Mileage (miles/(US)Gallon)")+
        ggtitle("Fitted model: mpg ~ TransmissionType*Weight")
print(g6)
```


Fitted model: mpg ~ TransmissionType*Weight
Gas Mileage (miles/(US)Gallon)
Weight (1000 lbs)
Manual
Auto
4 Cyl
6 Cyl
8 Cyl