A Study on Vehicles' Gas Mileage via Linear Regression Models

Wenhe 'Wayne' Ye

September 25, 2015

## Synopsis

## Data Processing

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

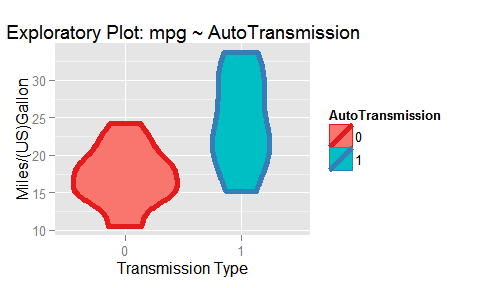
Clean up the raw data and convert some varibles into factors.

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 between mpg and factor of different transmission types to see the over all relationship.

(g1<-ggplot(data=mtcars2,aes(x=AutoTransmission,y=mpg,col=AutoTransmission))+  
 geom\_violin(aes(fill=AutoTransmission),size=2)  
+scale\_color\_brewer(palette="Set1")  
+ggtitle("Exploratory Plot: mpg ~ AutoTransmission")  
+labs(x="Transmission Type",y="Miles/(US)Gallon"))



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. The common sense tells us that most high-end vehicles, probably those gas guzzlers have automatic transmission rather than manual transmission. The later one is more likely to be found on some compact vehicles especially on the basic editions. 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 supporting information (SI)) It is worth noting here, we transform the cylinder number into factors rather than a continuous varible in the following study.

All the candidates showed some suspicious correlation with mpg. In order to quantify the difference between an auto transmission car and a manual transmission car, we need to carefully select the model to make our estimation.

## Regression Modeling

Our first attempt is to build a regression model includes all mentioned variables and factors. (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(fit\_all)

## GVIF Df GVIF^(1/(2\*Df))  
## disp 12.901490 1 3.591864  
## wt 6.821979 1 2.611892  
## hp 4.736101 1 2.176258  
## Cylinder 9.765272 2 1.767751  
## AutoTransmission 2.590898 1 1.609627

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, Cylinders and varible hp in the regression model. We also include ANOVA to verify if the variable inclusion is sufficient compare to *fit\_all* (in SI, P-value>0.05). Below is the coefficients of factors and varibles (A detailed table with hypothesis tests can be found in SI)

fit\_1<-lm(data=mtcars2,mpg~AutoTransmission+hp+Cylinder)  
coef(fit\_1)

## (Intercept) AutoTransmission1 hp Cylinder6   
## 27.29589929 4.15785647 -0.04424394 -3.92457850   
## Cylinder8   
## -3.53341392

The residual & diagnostics plot also suggests no obvious pattern existed in residual (in SI).

## Results

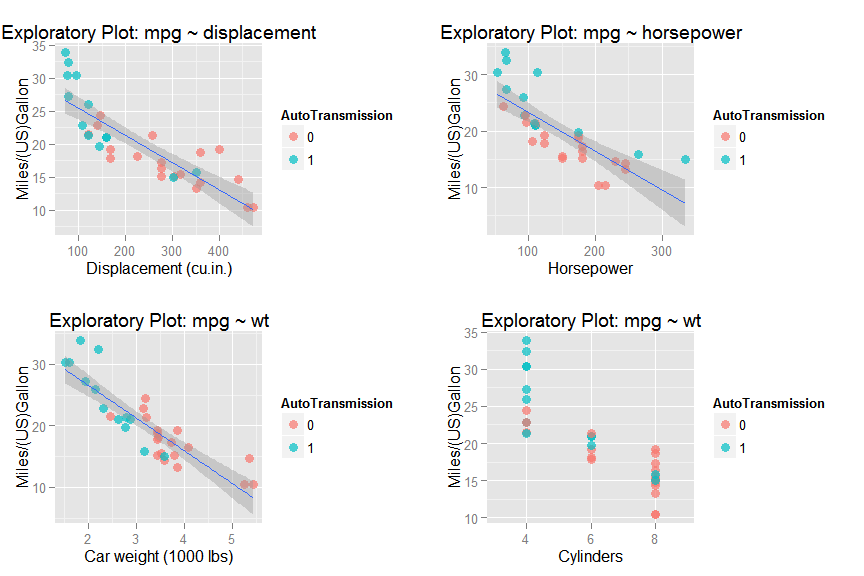
**Q1**

**Q2**

## *Supporting Information*

#### 1.mpg vs disp, hp, wt, Cylinder

g2<-ggplot(data=mtcars2,aes(x=disp,y=mpg))+geom\_point(size=3,alpha=0.7,aes(col=AutoTransmission))+  
 geom\_smooth(method="lm")+ggtitle("Exploratory Plot: mpg ~ displacement")+  
 labs(x="Displacement (cu.in.)",y="Miles/(US)Gallon")  
g3<-ggplot(data=mtcars2,aes(x=hp,y=mpg))+geom\_point(size=3,alpha=0.7,aes(col=AutoTransmission))+  
 geom\_smooth(method="lm")+ggtitle("Exploratory Plot: mpg ~ horsepower")+  
 labs(x="Horsepower",y="Miles/(US)Gallon")  
g4<-ggplot(data=mtcars2,aes(x=wt,y=mpg))+geom\_point(size=3,alpha=0.7,aes(col=AutoTransmission))+  
 geom\_smooth(method="lm")+ggtitle("Exploratory Plot: mpg ~ wt")+  
 labs(x="Car weight (1000 lbs)",y="Miles/(US)Gallon")  
g5<-ggplot(data=mtcars2,aes(x=Cylinder,y=mpg))+geom\_point(size=3,alpha=0.7,aes(col=AutoTransmission))+  
 ggtitle("Exploratory Plot: mpg ~ wt")+  
 labs(x="Cylinders",y="Miles/(US)Gallon")  
grid.arrange(g2,g3,g4,g5,ncol=2)



#### 2.ANOVA table

fit\_1<-lm(data=mtcars2,mpg~AutoTransmission+wt+Cylinder)  
fit\_all<-lm(data=mtcars2,mpg~.)  
anova(fit\_1,fit\_all)

## Analysis of Variance Table  
##   
## Model 1: mpg ~ AutoTransmission + wt + Cylinder  
## Model 2: mpg ~ disp + wt + hp + Cylinder + AutoTransmission  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 27 182.97   
## 2 25 150.41 2 32.56 2.7059 0.08634 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#### 3.Summary of model *fit\_1*

summary(fit\_1)

##   
## Call:  
## lm(formula = mpg ~ AutoTransmission + wt + Cylinder, data = mtcars2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.4898 -1.3116 -0.5039 1.4162 5.7758   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 33.7536 2.8135 11.997 2.5e-12 \*\*\*  
## AutoTransmission1 0.1501 1.3002 0.115 0.90895   
## wt -3.1496 0.9080 -3.469 0.00177 \*\*   
## Cylinder6 -4.2573 1.4112 -3.017 0.00551 \*\*   
## Cylinder8 -6.0791 1.6837 -3.611 0.00123 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.603 on 27 degrees of freedom  
## Multiple R-squared: 0.8375, Adjusted R-squared: 0.8134   
## F-statistic: 34.79 on 4 and 27 DF, p-value: 2.73e-10

#### 4.Residuals and Diagnostics

par(mfrow=c(2,2))  
for (i in 1:4){  
 plot(fit\_1,which=i)  
}

