A Study on Vehicles' Gas Mileage via Regression Models

Wenhe 'Wayne' Ye

September 25, 2015

Executive Summary

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 (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, Cylinders and varible wt in the regression model. The ANOVA table below indicates the variables included are sufficient compare to *fit_all* (P-value=0.5, which means we could not reject the null hypothesis).

```
## Analysis of Variance Table
##
## Model 1: mpg ~ AutoTransmission * wt + Cylinder
## Model 2: mpg ~ AutoTransmission * wt + Cylinder + disp + hp
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     26 137.99
## 2     24 130.38  2     7.615 0.7009  0.506
```

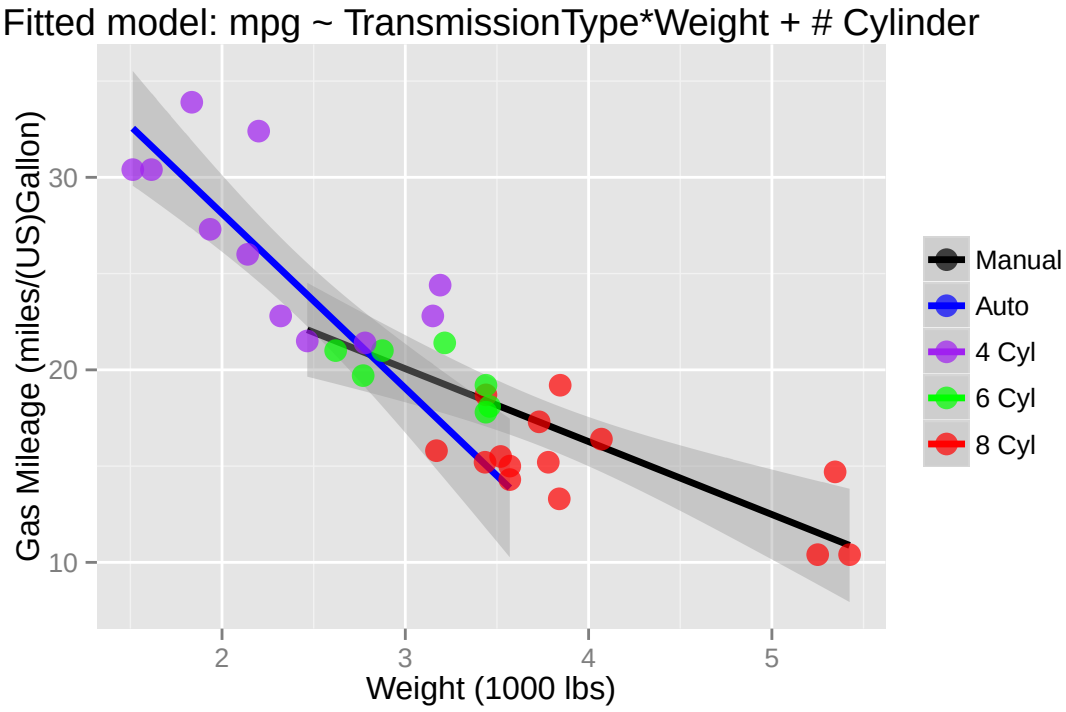
Below is the summary of the model we fit.

```
fit_1<-lm(data=mtcars2,mpg~AutoTransmission*wt+Cylinder)
summary(fit_1)
```

```
##
## Call:
## lm(formula = mpg ~ AutoTransmission * wt + Cylinder, data = mtcars2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5409 -1.5377 -0.6783  1.3160  5.2831
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             29.775      2.840  10.483 7.87e-11 ***
## AutoTransmission1       11.569      4.088   2.830  0.00885 **
## wt                      -2.399      0.844  -2.842  0.00860 **
## Cylinder6               -2.710      1.357  -1.996  0.05647 .
## Cylinder8               -4.776      1.556  -3.070  0.00496 **
## AutoTransmission1:wt    -4.068      1.397  -2.911  0.00730 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.304 on 26 degrees of freedom
## Multiple R-squared:  0.8775, Adjusted R-squared:  0.8539
## F-statistic: 37.23 on 5 and 26 DF,  p-value: 4.743e-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). Below is a plot for better interpreting:

```
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 + # Cylinder")
print(g6)
```



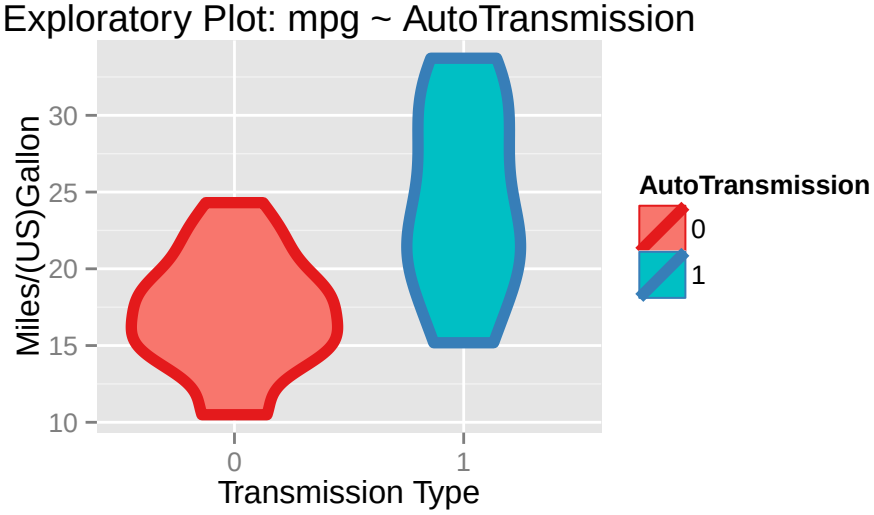
Results

Q1

Q2

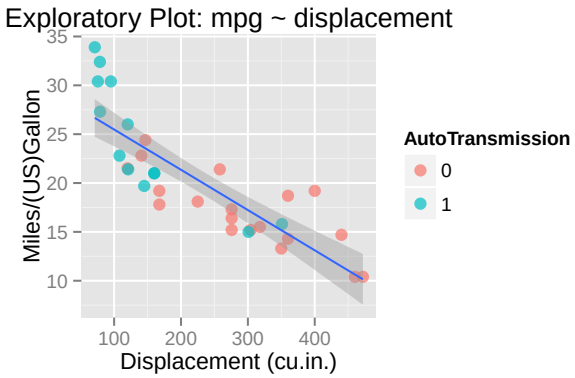
Supporting Information

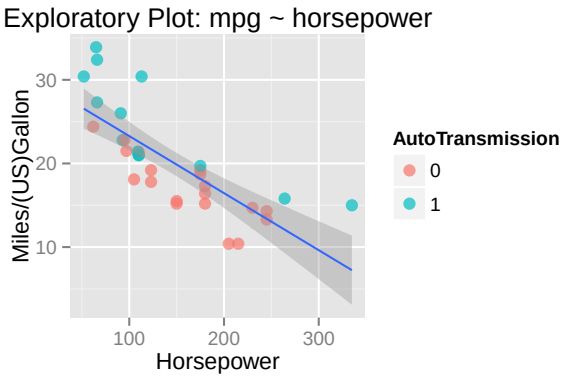
1.mpg vs Transmission Type Violin Plot

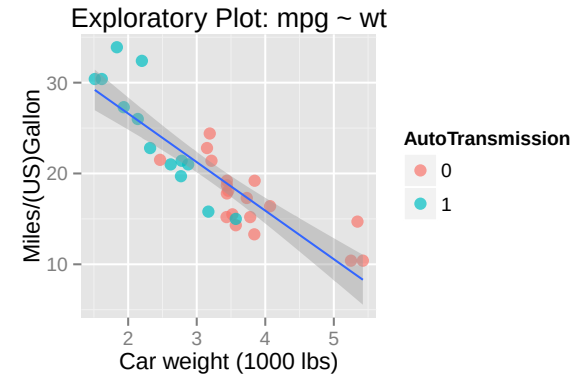


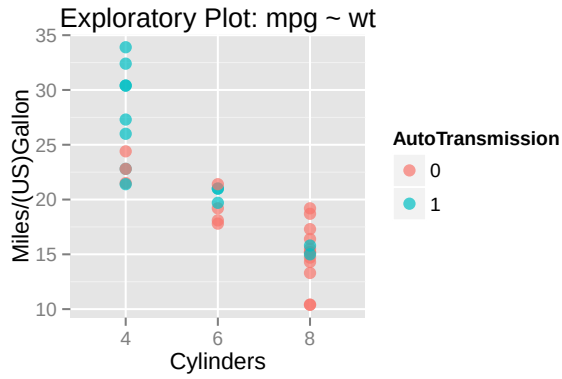
2.mpg vs disp, hp, wt,

Cylinder



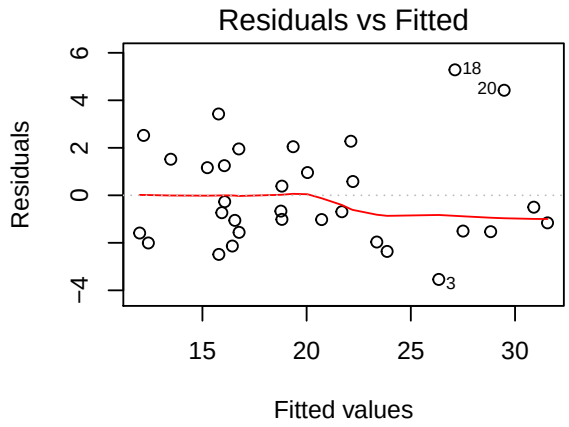


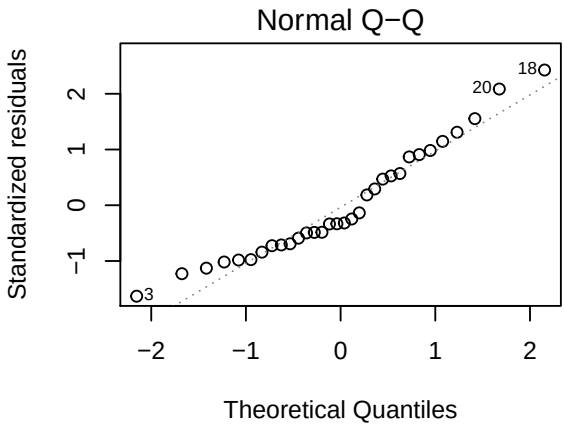


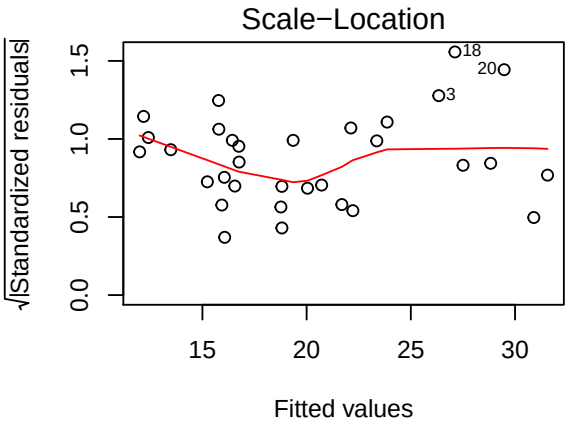


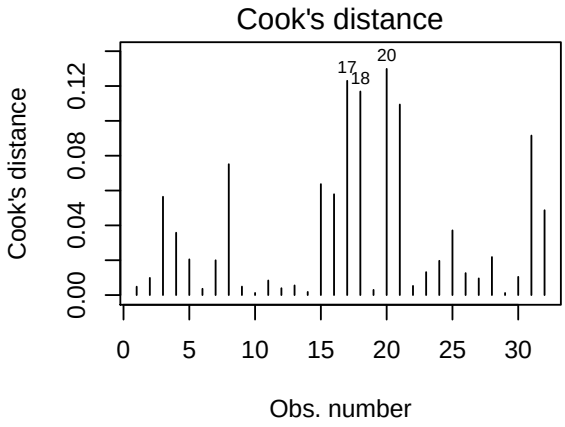
3.Residuals and Diagnostics

```
par(mfrow=c(2,2))
for (i in 1:4){
        plot(fit_1,which=i)
}
```


Residuals vs Fitted
Residuals
Fitted values
18
20
3


Normal Q−Q
Standardized residuals
Theoretical Quantiles
18
20
3


Scale−Location
√|Standardized residuals|
Fitted values
18
20
3


Cook's distance
Cook's distance
Obs. number
17
18
20