Automatic or Manual cars?

Becky

Saturday, July 26, 2014

Executive Summary

To understand the relationship between Miles per gallon (mpg) and Transmission, data of 32 car models are investigated. A regression model are thus developed. From the model, we can see the transmission has large effect on `mpg`. And as the coeffecient of the transmission large and negative, it indicates that manual cars can much lower the miles per gallon. And since the interactin of transmission with most of the other variables also have negative coeffcients in the model, the manual cars have less mile per gallon.


Synopsis

To understand the relationship between a set of variables and miles pre gallon (MPG), the data `mtcars` is used which is extracted from *Motor Trend* US magazine in 1974. The data comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles. In this report, we will particularly look at automatic or manual transmission is better for MPG, and quantify the MPG difference between them.

Exploratory Analysis

```r
data(mtcars)   #Load the data
#Preprocess the data
mtcars$am = factor(mtcars$am, levels=c(0,1), labels=c("automatic","manual"))
m=aggregate(mpg~am, data=mtcars, FUN=mean)   #mean

library(ggplot2)   #plot
qplot(,mpg, data=mtcars,color=am,
      xlab="Cars", ylab="Mile per Gallon (MPG)",
      main="MPG of different fuel consumption")+scale_colour_discrete(name="Transmission")
```
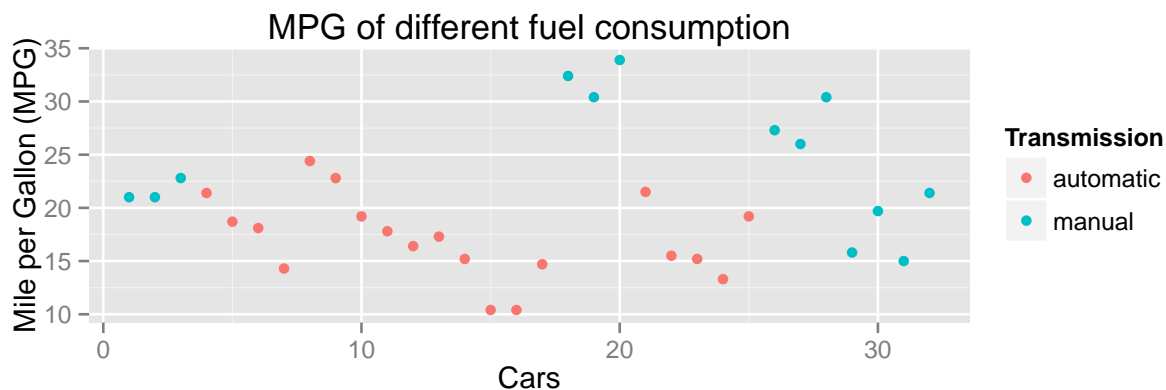

The mean mpg of **automatic** model is 17.1474 and that for the **manual** model is 24.3923. From the plot, we can notice that the manual model has higher mpg that that of the automatic model in general.

Regression Model

Model building

We first used the full dataset to build the model, and determine the importance of each variable by its p-value.

```
modelFull=lm(mpg~(.)*am, data=mtcars); summary(modelFull)
```

From the result, the p-value of each variable is high enough indicates we may undergo the model selection process for selecting important variables. *Summary of the model is shown in the appendix.*

Model selection

To find the best fit model, we then used the stepwise method to find a model with least AIC. The stepwise method will calculate the AIC for different combinations of variables and choose the model with least AIC. That is, the model with largest maximum likelihood and apporiate number of variables.

```
modelStep=step(modelFull, direction="both"); summary(modelStep)
```

The fitted model includes variables **displacement**, **Gross horsepower**, **Real axle ratio**, **weight**, **1/4 mile time**, **transmission**, **Number of forward gears**, and **Number of carburetors** to predict the **mpg**. *Summary of the model is shown in the appendix.*

Model Diagnostics

```
r= resid(modelStep)
par(mfrow=c(1,2))
plot(r, main="Residual plot", ylab="Residuals"); abline(h=0, col=4) #Residual plot
qqnorm(r); qqline(r,col=2)   #Normal-QQ plot
```
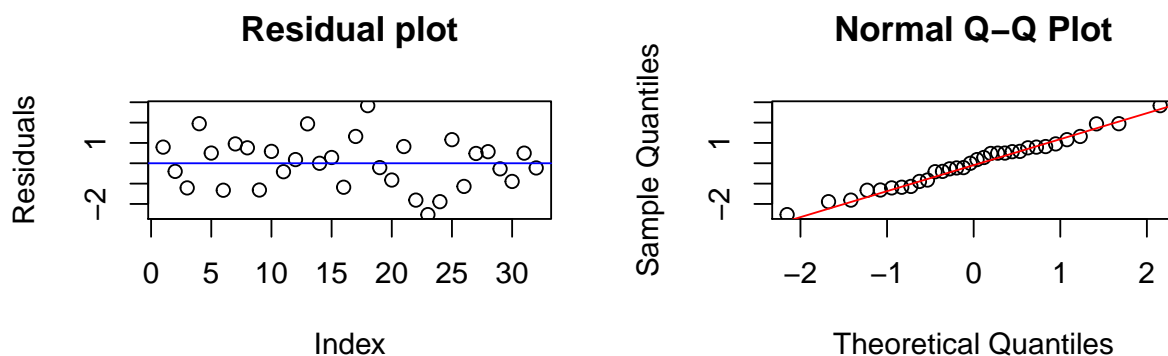

The plots indicates the resdiuals are around zero and scattered. The Normal QQ plot shows the model fit the normal distribution assumption.

Appendix

Summary of the full model

```
summary(modelFull)
```

```
##
## Call:
## lm(formula = mpg ~ (.) * am, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.035 -0.760  0.109  0.548  2.696
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     8.6435    22.3728    0.39    0.706
## cyl            -0.5339     1.1726   -0.46    0.657
## disp           -0.0203     0.0181   -1.12    0.286
## hp              0.0622     0.0479    1.30    0.218
## drat            0.5916     3.1326    0.19    0.853
## wt              1.9541     2.3207    0.84    0.416
## qsec           -0.8843     0.7888   -1.12    0.284
## vs              0.7389     2.6125    0.28    0.782
## ammanual     -146.5509    66.3235   -2.21    0.047 *
## gear            8.6542     4.0517    2.14    0.054 .
## carb           -4.8105     1.9765   -2.43    0.032 *
## cyl:ammanual   -0.7474     4.2614   -0.18    0.864
## disp:ammanual   0.2002     0.1596    1.25    0.234
## hp:ammanual    -0.2227     0.1381   -1.61    0.133
## drat:ammanual  -5.5414     5.8474   -0.95    0.362
## wt:ammanual   -12.4960     5.0728   -2.46    0.030 *
## qsec:ammanual   8.9793     3.2147    2.79    0.016 *
## vs:ammanual     0.2042     5.2854    0.04    0.970
## ammanual:gear   3.6743     7.2513    0.51    0.622
## ammanual:carb   9.4991     4.1683    2.28    0.042 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.88 on 12 degrees of freedom
## Multiple R-squared:  0.962,  Adjusted R-squared:  0.903
## F-statistic: 16.2 on 19 and 12 DF,  p-value: 8.25e-06
```

Summary of the fitted model

```r
summary(modelStep)
```

```
## 
## Call:
## lm(formula = mpg ~ disp + hp + drat + wt + qsec + am + gear +
##     carb + disp:am + hp:am + drat:am + wt:am + qsec:am + am:carb,
##     data = mtcars)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5233 -0.9544  0.0909  0.7692  2.8336
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -2.6791    10.7316   -0.25  0.80585
## disp           -0.0225     0.0148   -1.52  0.14613
## hp              0.0614     0.0300    2.04  0.05677 .
## drat            1.4504     1.9748    0.73  0.47269
## wt              2.1271     1.8887    1.13  0.27572
## qsec           -0.6665     0.5416   -1.23  0.23523
## ammanual     -103.6005    29.3389   -3.53  0.00256 **
## gear            9.3247     2.3836    3.91  0.00112 **
## carb           -5.1499     1.3613   -3.78  0.00148 **
## disp:ammanual   0.1330     0.0430    3.09  0.00662 **
## hp:ammanual    -0.1623     0.0578   -2.81  0.01209 *
## drat:ammanual  -4.4471     2.7375   -1.62  0.12266
## wt:ammanual   -11.7545     2.9142   -4.03  0.00086 ***
## qsec:ammanual   7.1815     1.5680    4.58  0.00027 ***
## ammanual:carb   8.2195     1.9222    4.28  0.00051 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.64 on 17 degrees of freedom
## Multiple R-squared:  0.959,  Adjusted R-squared:  0.926
## F-statistic: 28.6 on 14 and 17 DF,  p-value: 4.71e-09
```