# Regression Models Project Assignment

*Edie Hawthorne*

*7/18/2018*

# Overview

This is a report to analyze the relationship between a set of variables and miles per gallon(MPG). We are specifically interested in finding out whether is an automatic or manual transmission better for miles per gallon (MPG).

# Load the data

After loading the data and analyze the correlation among the set of variables in this dataset, we can see that mpg, cyl, wt disp are strongly correlated. So we are going to plug them in our models to analyze their residual variances.

```
library(dplyr)
library(ggplot2)
require(datasets)
require(GGally)
require(stats)
library(lmtest)

cor(mtcars)
```

```
##              mpg        cyl       disp         hp        drat         wt
## mpg   1.0000000 -0.8521620 -0.8475514 -0.7761684  0.68117191 -0.8676594
## cyl  -0.8521620  1.0000000  0.9020329  0.8324475 -0.69993811  0.7824958
## disp -0.8475514  0.9020329  1.0000000  0.7909486 -0.71021393  0.8879799
## hp   -0.7761684  0.8324475  0.7909486  1.0000000 -0.44875912  0.6587479
## drat  0.6811719 -0.6999381 -0.7102139 -0.4487591  1.00000000 -0.7124406
## wt   -0.8676594  0.7824958  0.8879799  0.6587479 -0.71244065  1.0000000
## qsec  0.4186840 -0.5912421 -0.4336979 -0.7082234  0.09120476 -0.1747159
## vs    0.6640389 -0.8108118 -0.7104159 -0.7230967  0.44027846 -0.5549157
## am    0.5998324 -0.5226070 -0.5912270 -0.2432043  0.71271113 -0.6924953
## gear  0.4802848 -0.4926866 -0.5555692 -0.1257043  0.69961013 -0.5832870
## carb -0.5509251  0.5269883  0.3949769  0.7498125 -0.09078980  0.4276059
##              qsec         vs          am       gear        carb
## mpg   0.41868403  0.6640389  0.59983243  0.4802848 -0.55092507
## cyl  -0.59124207 -0.8108118 -0.52260705 -0.4926866  0.52698829
## disp -0.43369788 -0.7104159 -0.59122704 -0.5555692  0.39497686
## hp   -0.70822339 -0.7230967 -0.24320426 -0.1257043  0.74981247
## drat  0.09120476  0.4402785  0.71271113  0.6996101 -0.09078980
## wt   -0.17471588 -0.5549157 -0.69249526 -0.5832870  0.42760594
## qsec  1.00000000  0.7445354 -0.22986086 -0.2126822 -0.65624923
## vs    0.74453544  1.0000000  0.16834512  0.2060233 -0.56960714
## am   -0.22986086  0.1683451  1.00000000  0.7940588  0.05753435
## gear -0.21268223  0.2060233  0.79405876  1.0000000  0.27407284
## carb -0.65624923 -0.5696071  0.05753435  0.2740728  1.00000000
```

# Model Selection

First we plug in all of the variables into our model (fit) comparing with just plug in one variable of cyl in our model (fit1), from the anova table we can see that including one variable is more significant than including all variables in our model according to our P value. Then we plug in one more variable of weight (wt) in to a new model (fit3), we can see that including one variables of cyl and wt is more significant than including just cyl according to our P value. Then we plug in the disp variable into fit2 model, we can see it become less significant comparing with the fit2 model by just including cyl and wt in our model.

From the analysis of the likelihood ratio, we also see that the model fit2 which includes the cyl and wt in has the least log likelihood compared with the other models.

From this analysis, we decided to move on with the fitted model of fit2 to continue our analysis.

```
fit <- lm(mpg~., mtcars)
fit1 <- lm(mpg~factor(am) + factor(cyl), mtcars)
fit2 <- lm(mpg~factor(am) + factor(cyl) + wt, mtcars)
fit3 <- lm(mpg~factor(am) + factor(cyl) + wt + disp, mtcars)
null_fit <- lm(mpg~1, mtcars)

anova(null_fit, fit, fit1, fit2, fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ 1
## Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## Model 3: mpg ~ factor(am) + factor(cyl)
## Model 4: mpg ~ factor(am) + factor(cyl) + wt
## Model 5: mpg ~ factor(am) + factor(cyl) + wt + disp
##   Res.Df     RSS Df Sum of Sq       F    Pr(>F)
## 1     31 1126.05
## 2     21  147.49 10    978.55 13.9325 3.793e-07 ***
## 3     28  264.50 -7   -117.00  2.3798  0.058622 .
## 4     27  182.97  1     81.53 11.6077  0.002654 **
## 5     26  182.87  1      0.10  0.0141  0.906621
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lrtest(null_fit, fit, fit1, fit2, fit3)
```

```
## Likelihood ratio test
##
## Model 1: mpg ~ 1
## Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## Model 3: mpg ~ factor(am) + factor(cyl)
## Model 4: mpg ~ factor(am) + factor(cyl) + wt
## Model 5: mpg ~ factor(am) + factor(cyl) + wt + disp
##   #Df   LogLik Df   Chisq Pr(>Chisq)
## 1   2 -102.378
## 2  12  -69.855 10 65.0457  3.973e-10 ***
## 3   5  -79.199 -7 18.6891  0.0092191 **
## 4   6  -73.303  1 11.7924  0.0005947 ***
## 5   7  -73.295  1  0.0173  0.8952971
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Fitted model

By looking at the summary of the fitted model, using automatics transmission as reference and holding all the other variables at their mean values, switching to the manual transmission will increase the mile per gallon 0.15.

```
fit2 <- lm(mpg~factor(am) + factor(cyl) + wt, mtcars)
summary(fit2)$coef
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  33.7535920  2.8134831 11.9970836 2.495549e-12
## factor(am)1   0.1501031  1.3002231  0.1154441 9.089474e-01
## factor(cyl)6 -4.2573185  1.4112394 -3.0167231 5.514697e-03
## factor(cyl)8 -6.0791189  1.6837131 -3.6105432 1.227964e-03
## wt           -3.1495978  0.9080495 -3.4685309 1.770987e-03
```

But based on the P value for the tranbsmission variable, we fail to reject the null hypothesis that there is a difference in miles per gallon between automatic and manual transmission. This can also be seen in the confidence interval that the confidence interval of the diference in automatical and manual transmission includes 0, which means we fail to reject null hypothesis (which is assums that there is no difference between the automatic and manual transmissions).

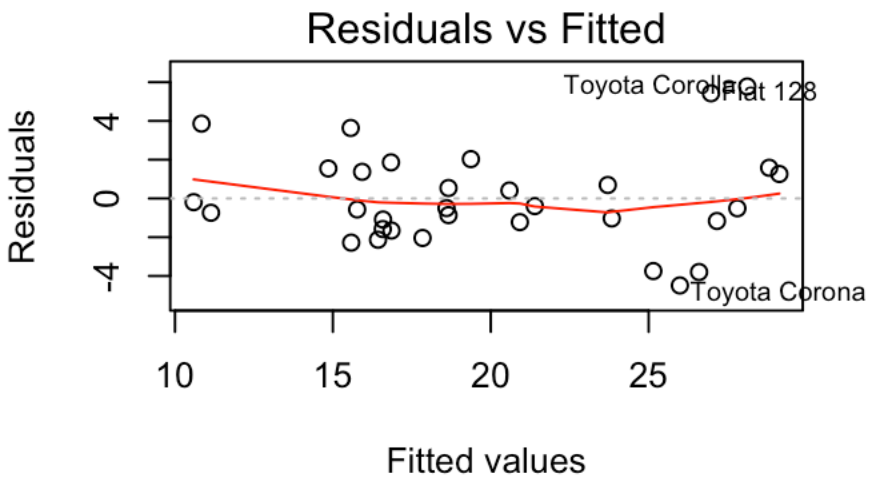
```
conf_interval <- 0.15 + c(-1,1)*qnorm(0.975)*1.3/sqrt(32)
```

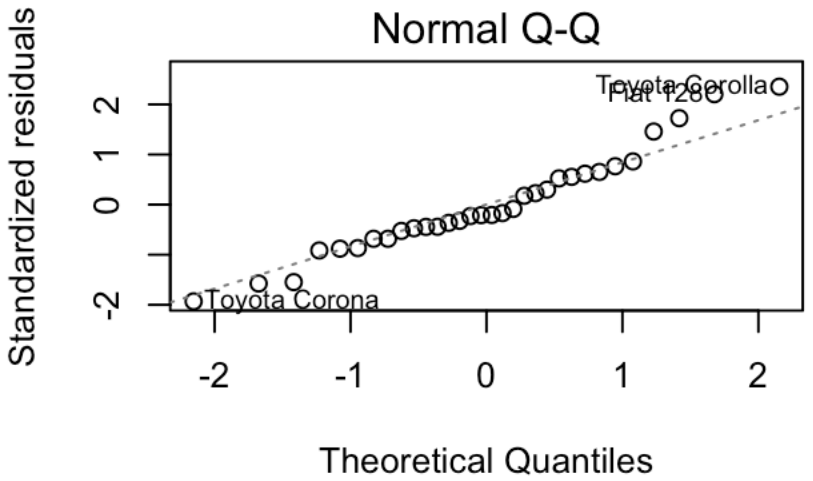
Confidence interval for transmission: -0.30, 0.60

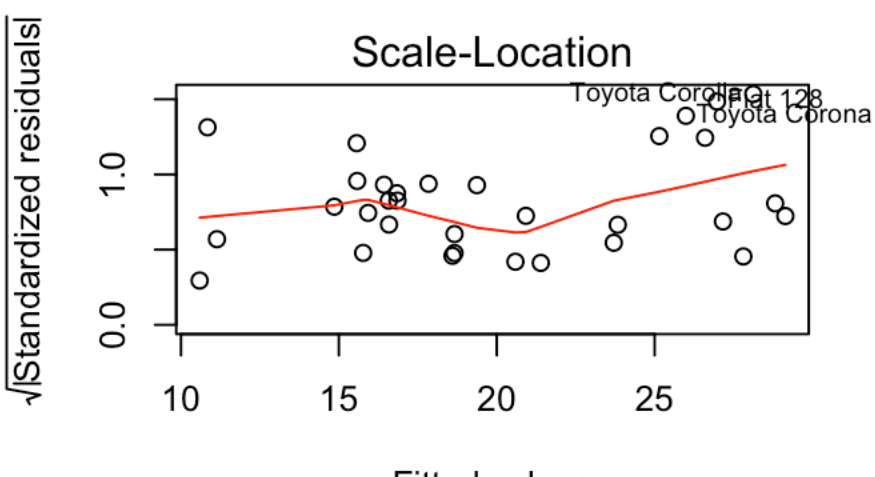
# Residual Diagnostics and Variations

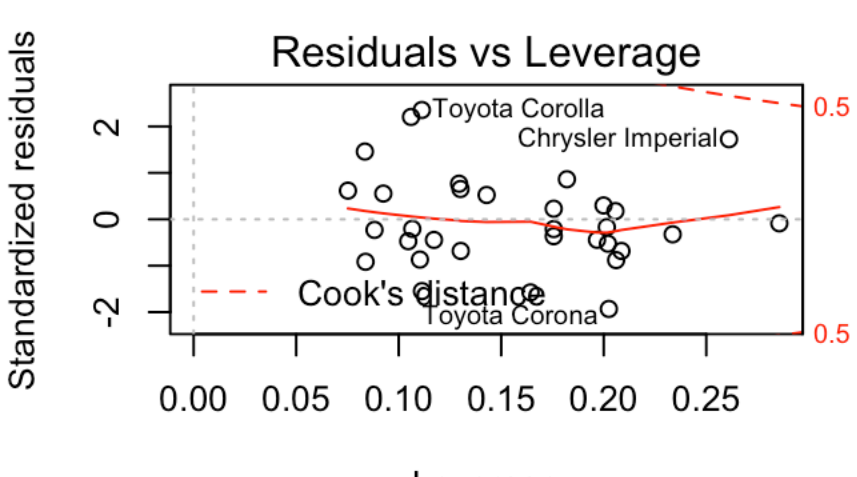
By ploting the residual analysis, the assumptions for the error term of constant variance, indenpendence, and normality are met.

```
par(mfrow=c(2,2))
plot(fit2)
```




# Conclusion

From our analysis, we find that there is no statistical difference between automatical and manual transmissions.