Regression_final

This project aims to explore: 1) whether an automatic or manual transmission is better for MPG; 2) the difference between automatic and manual transmissions on the MPG.

The work is for Motor Trend, a magazine about the automobile industry.

1. Setup workspace and load packages.

2. Explore data

```
raw_data<-mtcars

#explore variables

str(raw_data)
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...

summary(raw_data)
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
##        am              gear            carb      
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
##  Median :0.0000   Median :4.000   Median :2.000  
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812  
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

3. Run regression models

Model 1: mpg~am; Model 2: mpg~am+cyl+wt+hp; Model 3: mpg~am+wt+hp

Notes: I am not an expert on analyzing automobile data.With my limited knowledge, "wt", "hp", "cyl" seem to be possible predictors for mpg. So in the model 2, I try to control for those variables.

```
## change am to categorical variable (1: auto; 0:manual)
raw_data$am<-as.character(raw_data$am)

#model 1: mpg~am
```

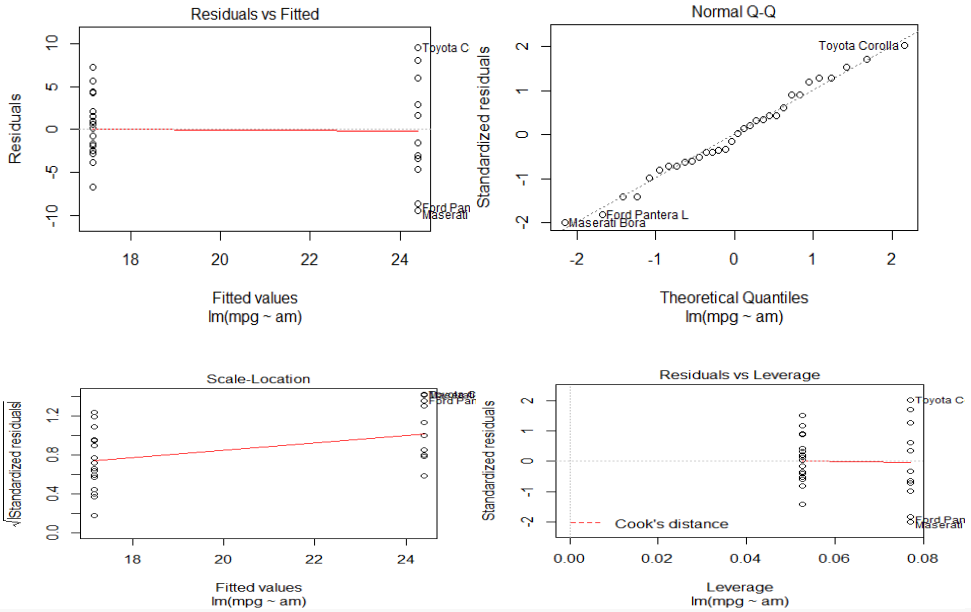
```
model1<-lm(mpg~am, raw_data)
summary(model1)$coef
##             Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am1          7.244939   1.764422  4.106127 2.850207e-04
# Result for Model1: The result of linear regression shows that automatic transmission is significantly better than manual transmission for MPG. The MPG difference between automatic and manual transmissions is 7.24 (p<.001)

# Model 2
raw_data$cyl<-as.character(raw_data$cyl)
Model2<-lm(mpg~am+cyl+wt+hp, raw_data)
summary(Model2)$coef
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 33.70832390 2.60488618 12.940421 7.733392e-13
## am1          1.80921138 1.39630450  1.295714 2.064597e-01
## cyl6        -3.03134449 1.40728351 -2.154040 4.068272e-02
## cyl8        -2.16367532 2.28425172 -0.947214 3.522509e-01
## wt          -2.49682942 0.88558779 -2.819404 9.081408e-03
## hp          -0.03210943 0.01369257 -2.345025 2.693461e-02
# Result for Model2: The result of linear regression shows that automatic transmission is NOT significantly better than manual transmission for MPG after controlling for cyl (P=0.21).

# Model 3
Model3<-lm(mpg~am+wt+hp, raw_data)
summary(Model3)$coef
##               Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 34.00287512 2.642659337 12.866916 2.824030e-13
## am1          2.08371013 1.376420152  1.513862 1.412682e-01
## wt          -2.87857541 0.904970538 -3.180850 3.574031e-03
## hp          -0.03747873 0.009605422 -3.901830 5.464023e-04
# Result for Model3: The result of linear regression shows that automatic tranmission is NOT significantly better than manual transmission for MPG after controlling for cyl (P=0.14).The MPG difference is 2.08

#compare model 1,model 2 and model3 by using ANOVA
anova(model1, Model2, Model3)
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl + wt + hp
## Model 3: mpg ~ am + wt + hp
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.5267 1.688e-08 ***
## 3     28 180.29 -2    -29.27  2.5191       0.1 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
# The result of ANOVA shows that Model 2 fit the data significantly better than model 1 (P<.001), Model 2 and Model 3 are not significantly different. So Model 3 is the best model.
```

In summary, the results of this project indicate an automatic transmission is not better than a manual transmission for MPG. According to the model comparsion, the Model 3 is the best model. It shows the MPG difference between automatic and manual transmission is 2.08, but not significantly different from zero (p=0.14).

Appendix: Supporting Residual plots

#1. Review residual plots of the model 1

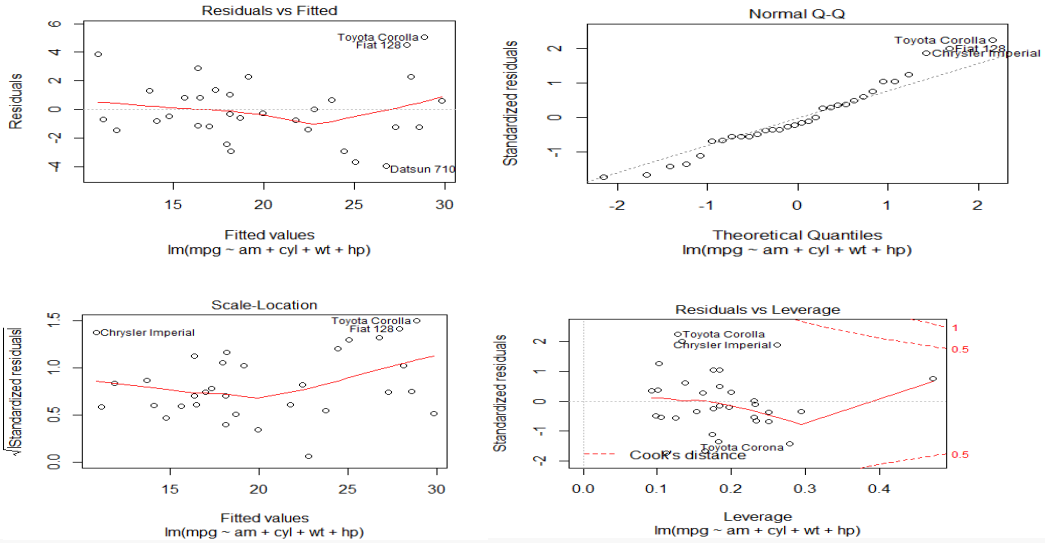
plot(model1)



 # All plots seem not too bad, although the Scale_location plot shows that standarized residuals slightly increase with the fitted values.

#2. review residule plots of the model 2

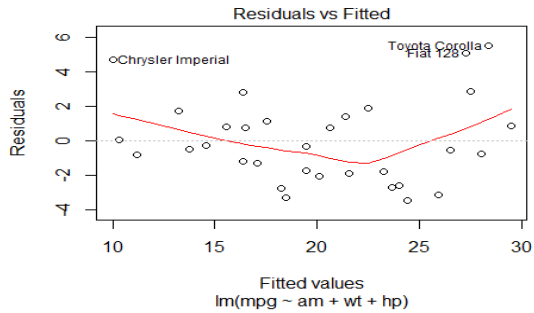
plot(Model2)

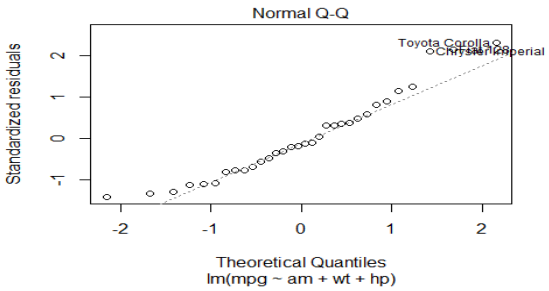


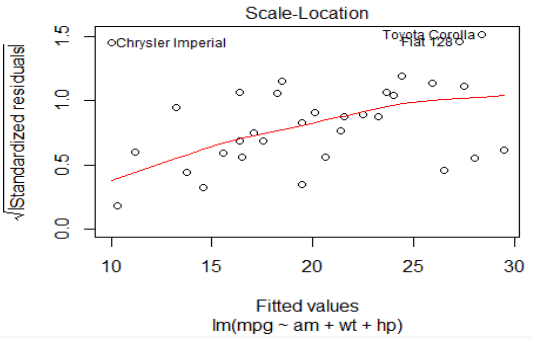
 # All plots seem not too bad, although the ResidualsvsFitted and Scale_location plots show that (standarized) residuals slightly increase with the fitted values.

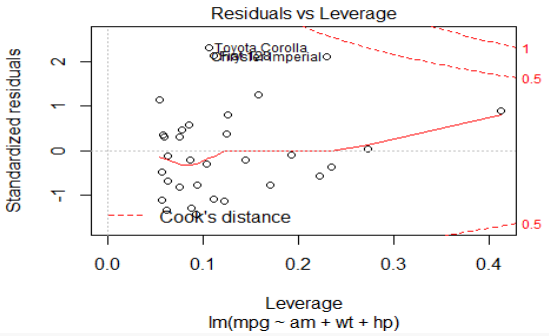
#3. review residule plots of the model 3

plot(Model3)









 # All plots seem not too bad, although the ResidualsvsFitted and Scale_location plots show that (standarized) residuals slightly increase with the fitted values.