Regression Model PA - Jingchen LI

Wednesday, August 20, 2014

- Question 1: which transmission is better for MPG (miles/gallon)? Automatic or manual?
- Question 2: quantify the MPG difference between automatic and manual transmissions.

Summary

I adopt the model:$log(mpg) = \beta_0 + \beta_1 wt + \beta_2 vs + \beta_3 am + \beta_4 wt{:}\, am + e$

The model answers two question: 1. Manual transmissions are better at MPG and this betterment is interacted with weight of the car, the lighter the car, the larger the betterment. 2. The model predicts, if given the same weight, let's say w, and same V/S, a manual car multiple the MPG performance over an auto car, by e^(0.36412-0.12468w).

The model has a balance between credibility and explainability. And it's relatively simple as having one continuous predictor, two categorical predictor and one interaction term on the right and logized mpg on the left.

Finally, I am not a car lover and not even a driver, so I have no knowledge about the car. If you find this model contradicting some "common sense" about car, please let me know!

Explaination of the findings; reproduciable codes in the appendix

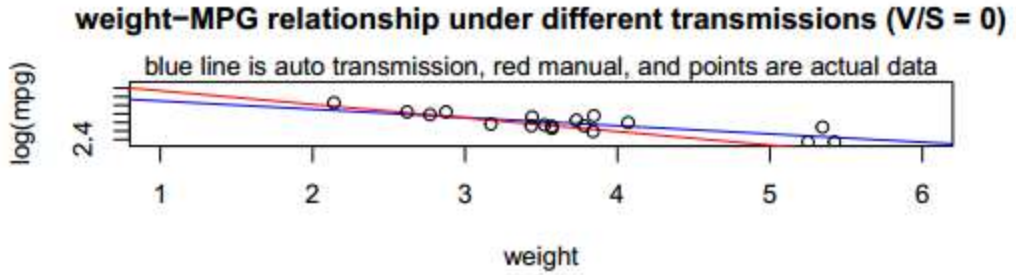
The model: $log(mpg) = \beta_0 + \beta_1 wt + \beta_2 vs + \beta_3 am + \beta_4 wt{:}\, am + e$

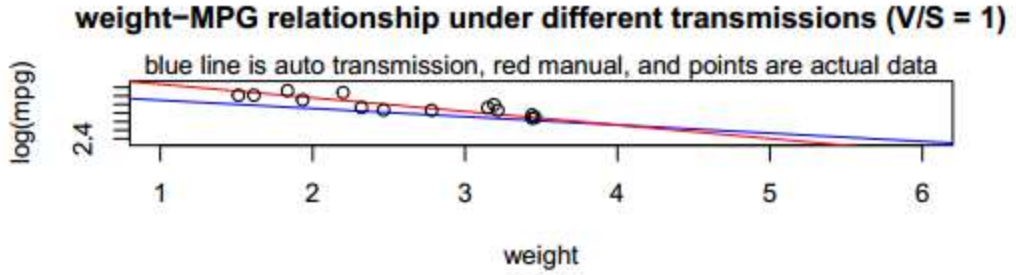
It identifies three predictors: wt-weight of the car, measured in per 1000 lbs; vs - V engine or straight engine, "0" is v engine and "1" is straight engine, and am - automatic or manual transmission, "0" is auto and "1" is manual.

The model takes 4 steps to carry out: 1. Regress mpg on all possible single variables in the `mtcars` dataset. 2. Delete unneccessary variables due to confounding issue; 3. Add interaction terms based on simplified model got in step 2; 3. Check linearity, normality and outliers. Finally use `log(mpg)` instead of `mpg` as outcome to have better normality and keep good credibility and explainability

Details of the birth of this model is in the appendix.

The model finds two relationships from transmission. First, a manual transmission are related to higher MPG. The merely change from auto to manual will maginify MPG by around e^0.36; second, this magnification will shrink (multiple) by around e^(-0.125) for every 1000 lbs car weights. Therefore, under different transmissions, MPG has a different realtionship between weight. The plots might show this relationship better.





The plot predicts that weight decrease the MPG, but auto and manual transimission cars are having different rate of MPG decreasing; manual ones decrease faster. **when having same V/S**,before weight around 3000 lbs, manual transmission car have a higher MPG but after that point, auto transmission cars become more fuel efficient.

Besides, R square and adjusted R square of this model are high enough.

```
## $`R2 and adjusted R2`
## [1] 0.8374
```

The model has sufficient confidence as its P-values are low.

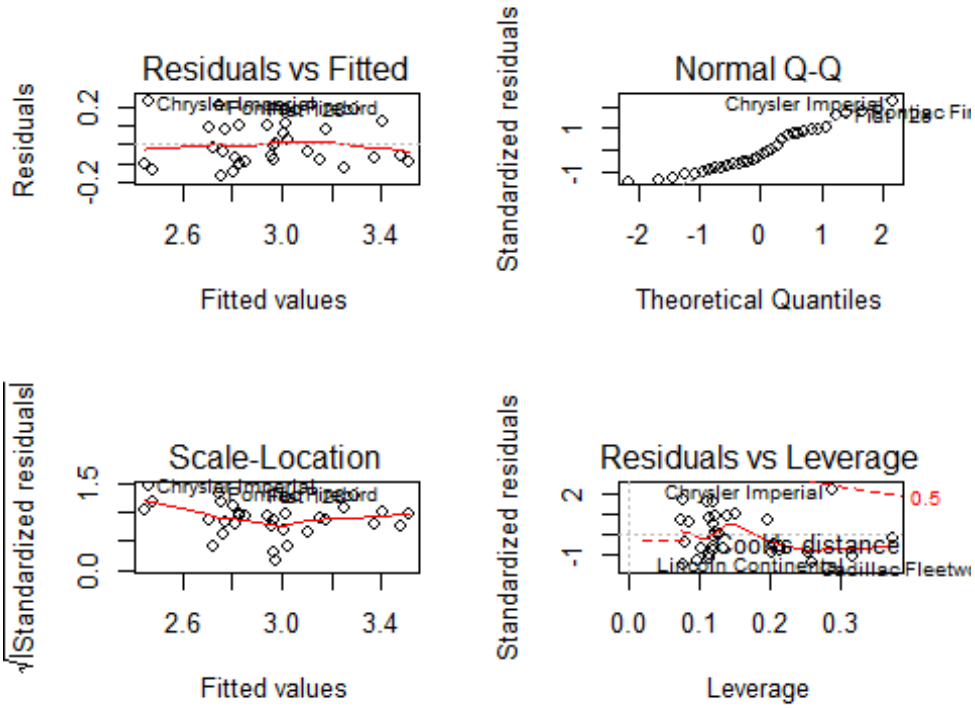
```
## (Intercept)          wt         vs1         am1      wt:am1
##   6.359e-18   9.551e-05   2.315e-02   7.675e-02   7.791e-02
```

Although P-values for coefficients of `am1` and `wt:am1` are around 8%, given only 32 observations, I deem this is a sufficient level of confidence.

Although the QQ plot doesn't seems to be good enough. In fact, I already taken log of mpg is to reduce the distortion on normality. What is presented here is much better then the previous model: $mpg = \beta_0 + \beta_1 wt + \beta_2 vs + \beta_3 am + \beta_4 wt{:}\, am + e$. The large deviance from normality indicates that the data itself isn't normal.

For the outliers presented in the plots, I have checked their nature and decided to keep them because 1. the outliers are usual home cars; 2. if delete them from the data, the normality get worse and new outliers coming out, implying those "outliers" are just part of this dataset.

The model also has satisfactory linearity and normality as shown by following plots.



Things to improve

First, the normality of this model is great, perhaps new methods can be applied but I don't know.

Second, an alternative model without the `vs` predictor $mpg = \beta_0 + \beta_1 wt + \beta_2 am + \beta_3 wt{:}\, am + e$ It actually has lower P-values. However, I include `vs` because it does take a plcae in explaining mpg even though it's slightly associated with `am` and therefore decreases P-lvaues of current model. I am not sure about the trade-off between them.

Last and perhaps most important, I don't have enough car knowedge to examine this model in terms of real meaning. The selection of variables, interactions and trade-off between explainabilit and credibility are all effected by my lack of car knowedge. Hopefully, I will know more and make a better model when I drive a car in a year or so.

Appendix

Here you can reproduce all data and model and review my thoughts

load data

```
data(mtcars)
mt <- mtcars
summary(mt)
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.4   Min.   :4.00   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.4   1st Qu.:4.00   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.2   Median :6.00   Median :196.3   Median :123.0  
##  Mean   :20.1   Mean   :6.19   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.8   3rd Qu.:8.00   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.9   Max.   :8.00   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.76   Min.   :1.51   Min.   :14.5   Min.   :0.000  
##  1st Qu.:3.08   1st Qu.:2.58   1st Qu.:16.9   1st Qu.:0.000  
##  Median :3.69   Median :3.33   Median :17.7   Median :0.000  
##  Mean   :3.60   Mean   :3.22   Mean   :17.8   Mean   :0.438  
##  3rd Qu.:3.92   3rd Qu.:3.61   3rd Qu.:18.9   3rd Qu.:1.000  
##  Max.   :4.93   Max.   :5.42   Max.   :22.9   Max.   :1.000  
##        am              gear            carb      
##  Min.   :0.000   Min.   :3.00   Min.   :1.00  
##  1st Qu.:0.000   1st Qu.:3.00   1st Qu.:2.00  
##  Median :0.000   Median :4.00   Median :2.00  
##  Mean   :0.406   Mean   :3.69   Mean   :2.81  
##  3rd Qu.:1.000   3rd Qu.:4.00   3rd Qu.:4.00  
##  Max.   :1.000   Max.   :5.00   Max.   :8.00
```

There are some categorical variables needed to be factorized otherwise will cause trouble in R

```
mt$cyl <- as.factor(mt$cyl)
mt$vs <- as.factor(mt$vs)
mt$am <- as.factor(mt$am)
mt$gear <- as.factor(mt$gear)
mt$carb <- as.factor(mt$carb)
```

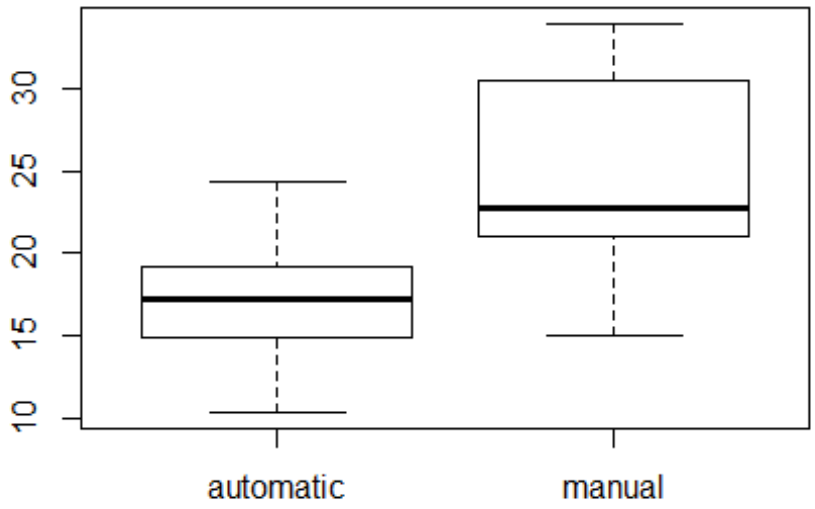
This is a categorical variable, first make a comparison on average value:

```
avgauto <- mean(mt$mpg[mt$am==0])
avgmanual <- mean(mt$mpg[mt$am==1])
print(data.frame(avgauto,avgmanual))
```

```
##   avgauto avgmanual
## 1   17.15     24.39
```

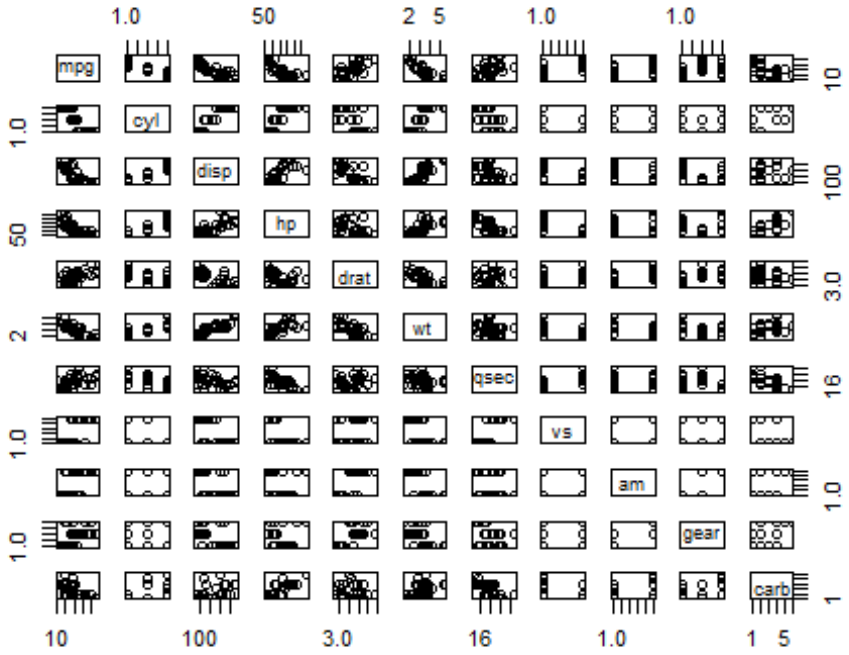
Rough feeling is that manual is better.Look deepr into the distribution of mpg for two categories:

```
boxplot(data=mt,mpg~am,names=c("automatic","manual"))
```



Seems the distribution of the two mpgs are also normal. More tendency that manual is better. But! We are professional, so we need to consider the possible confounding effects from other variables. Let's make a scatterplot matrix of all variables in the mtcars dataset.

```
pairs(mt)
```



Each bivariate relationship seems to have a pattern.The plot show a high correlation between variables.So there is porbably confounding effects in the relationship between "mpg" and "cyl". A multivariable regression is needed to control the influence from other variables to mpg.

Since there are categorical variables, we need to notify R about this to prvent it treats every independent variable as continuouse.

```
model1 <- lm(data=mt, mpg~.)
summary(model1)$coef
```

```
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 23.87913   20.06582  1.19004  0.25253
## cyl6        -2.64870    3.04089 -0.87103  0.39747
## cyl8        -0.33616    7.15954 -0.04695  0.96317
## disp         0.03555    0.03190  1.11433  0.28267
## hp          -0.07051    0.03943 -1.78835  0.09393
## drat         1.18283    2.48348  0.47628  0.64074
## wt          -4.52978    2.53875 -1.78426  0.09462
## qsec         0.36784    0.93540  0.39325  0.69967
## vs1          1.93085    2.87126  0.67248  0.51151
## am1          1.21212    3.21355  0.37719  0.71132
## gear4        1.11435    3.79952  0.29329  0.77332
## gear5        2.52840    3.73636  0.67670  0.50890
## carb2       -0.97935    2.31797 -0.42250  0.67865
## carb3        2.99964    4.29355  0.69864  0.49547
## carb4        1.09142    4.44962  0.24528  0.80956
```

```
## carb6        4.47757    6.38406  0.70137  0.49381
## carb8        7.25041    8.36057  0.86722  0.39948
```

All variables are having high P-values, indicating the model are having too much confounding variables. So the next step is too deleting the unneccessary variables.Normally, this process should start by deleting variabls with highest P-value. However, if the variable is categorical, we can't technically delete a value of this vairable. So the deleting process will apply only to continuose variables.

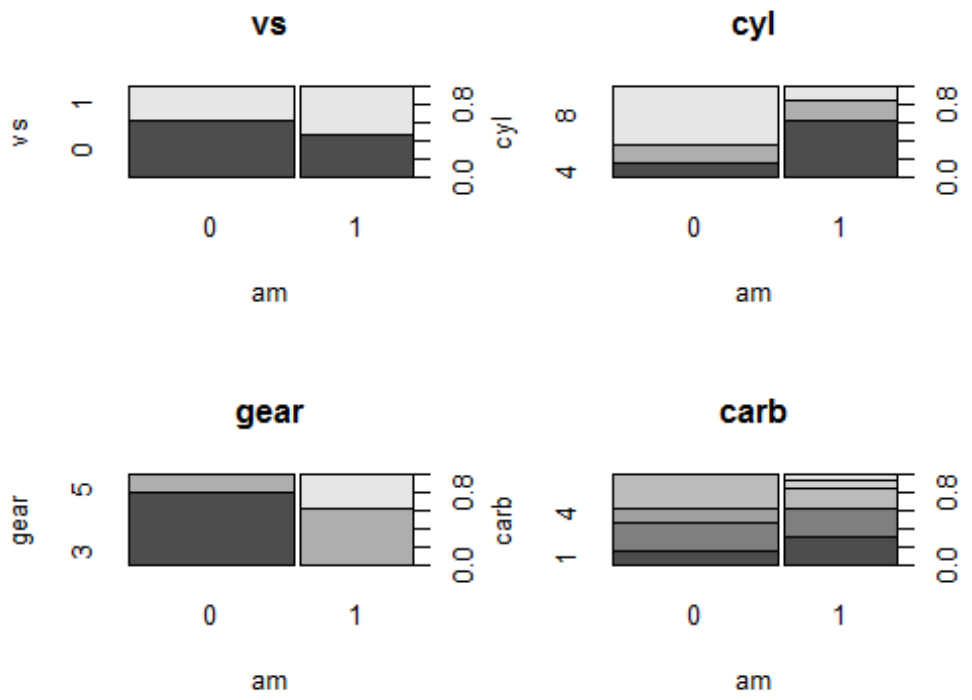
The variable "drat", "disp" and "qsec" has high P-values and they are continuouse so let's start by deleting them

```
model2 <- update(model1,.~.-qsec)
model3 <- update(model2,.~.-drat)
model4 <- update(model3,.~.-disp)
model5 <- update(model4,.~.-hp)
model6 <- update(model5,.~.-wt)
```

In model4, the continuous variable "hp" and "wt" are having P-values around 10%, which is not perfect but enough considering we only have 32 observations. Further deletion of any of them will led to significant drop in R square and adjuested R square. Therefore, I will keep them in the model for now and consider model4 as the step stone for next investigation. Now, the focus is turned to categorical variables left in the model.

Since the P-values of all categorical variables are very high and previous pairs plot show that correaltion between these variables are also high. Confounding issues could be disrupting. We need to reduce confounding effect on "am". Let's see how those categorical variables are confunded by regress each of them on "am" by making plots against "am".

```
par(mfrow=c(2,2))
plot(vs~am,data=mt,main=c("vs"))
plot(cyl~am,data=mt,main=c("cyl"))
plot(gear~am,data=mt,main=c("gear"))
plot(carb~am,data=mt,main=c("carb"))
```



The plots show that those categorical variables are far from randomly distributed when am=0 and am=1. Theresfore, strong confounding could exist. In this case, we need to delete categorical variables that are likely to be associated with "am". The plots show that "gear" and "cyl" has very different distribution between am=0 and am=1, so let's start from deleting them.

```
model5 <- update(model4, .~.-gear-cyl)
summary(model5)

##
## Call:
## lm(formula = mpg ~ hp + wt + vs + am + carb, data = mt)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.297 -1.287  0.031  0.760  5.243
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  29.0272     4.7216    6.15  3.5e-06 ***
## hp           -0.0266     0.0161   -1.66    0.112
## wt           -2.0735     1.0728   -1.93    0.066 .
## vs1           1.9041     1.7239    1.10    0.281
## am1           3.2593     1.7188    1.90    0.071 .
## carb2         0.1710     1.4844    0.12    0.909
## carb3         0.0651     2.3211    0.03    0.978
## carb4        -1.5397     1.8669   -0.82    0.418
```

```
## carb6        -2.1873     3.2425   -0.67    0.507
## carb8        -0.9720     4.4593   -0.22    0.829
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.66 on 22 degrees of freedom
## Multiple R-squared:  0.861,  Adjusted R-squared:  0.805
## F-statistic: 15.2 on 9 and 22 DF,  p-value: 1.63e-07
```

The adjusted R2 has improved but P-values are still far from crediable, especially for "carb". So delete it.

```
model6 <- update(model5, .~.-carb)
summary(model6)
```

```
##
## Call:
## lm(formula = mpg ~ hp + wt + vs + am, data = mt)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.671 -1.788 -0.304  1.289  5.330
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  31.0788     3.3928    9.16    9e-10 ***
## hp           -0.0301     0.0109   -2.75   0.0105 *
## wt           -2.5910     0.9174   -2.82   0.0088 **
## vs1           1.7855     1.3271    1.35   0.1897
## am1           2.4171     1.3794    1.75   0.0911 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.5 on 27 degrees of freedom
## Multiple R-squared:  0.85,   Adjusted R-squared:  0.828
## F-statistic: 38.2 on 4 and 27 DF,  p-value: 9.45e-11
```

The adjusted R2 has improved but P-values are crediable, except for "vs". So delete it.

```
model7 <- update(model6, .~.-vs)
summary(model7)
```

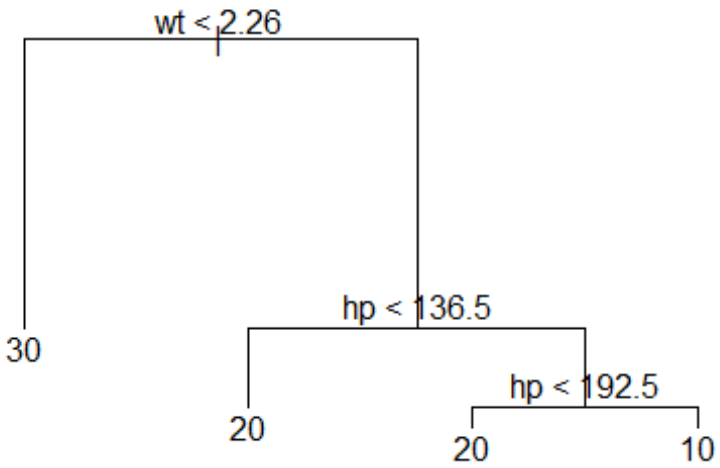
```
##
## Call:
## lm(formula = mpg ~ hp + wt + am, data = mt)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.422 -1.792 -0.379  1.225  5.532
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.00288    2.64266   12.87  2.8e-13 ***
## hp          -0.03748    0.00961   -3.90  0.00055 ***
## wt          -2.87858    0.90497   -3.18  0.00357 **
## am1          2.08371    1.37642    1.51  0.14127
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.54 on 28 degrees of freedom
## Multiple R-squared:  0.84,   Adjusted R-squared:  0.823
## F-statistic:   49 on 3 and 28 DF,  p-value: 2.91e-11
```

The adjusted R2 hasn't changed but P-value for am decrease, indicating vs could have its place in the model. SO let's go back to model6.

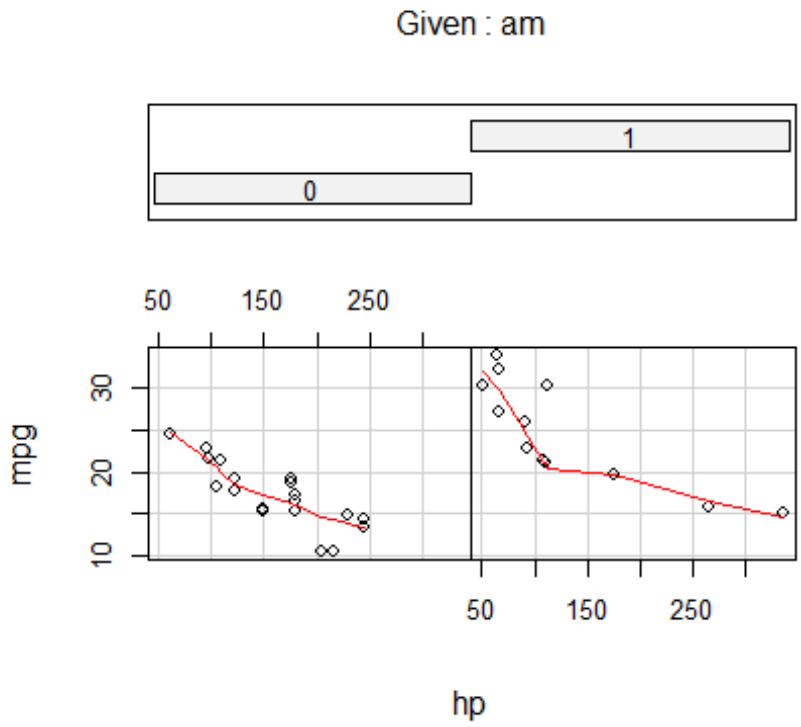
Now we only has 4 predictors and only "vs" has a P-value bigger than 10%, the line I deemed as credible, according to the size of 32 observations. Let's make a tree to see what could possible be the combination of these 4 variabls to influence "mpg"

```
require(tree)

## Loading required package: tree

model6.tree <- tree(mpg~hp+wt+am+vs,data=mt)
par(mfrow=c(1,1))
plot(model6.tree)
text(model6.tree)
```

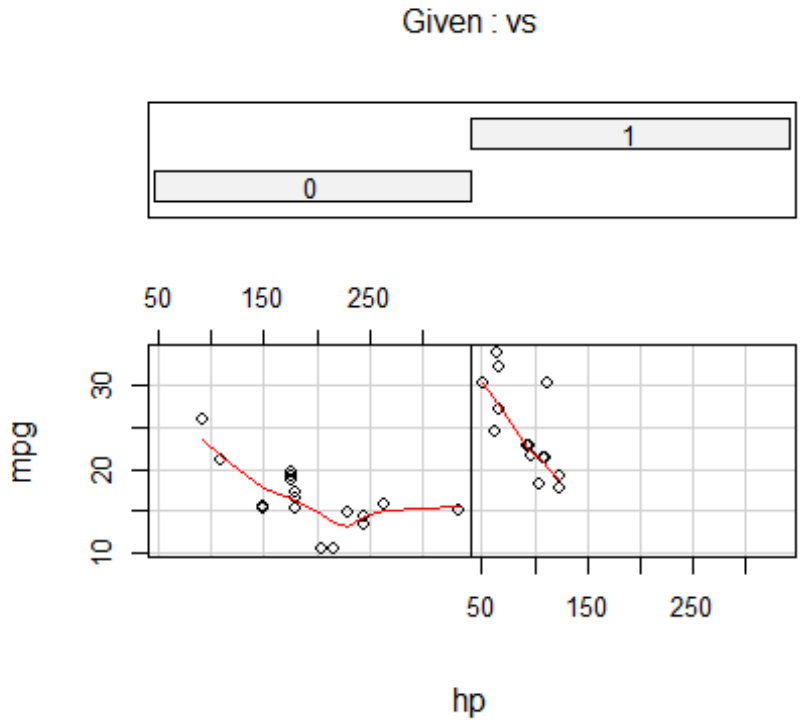


UNfortunately, binary tree doesn't give me information about the influence of categorical variables "am" and "vs", meaning they are not that influential on their own. However, there might be interaction between variabls that makes "am" and "vs" insignificant.

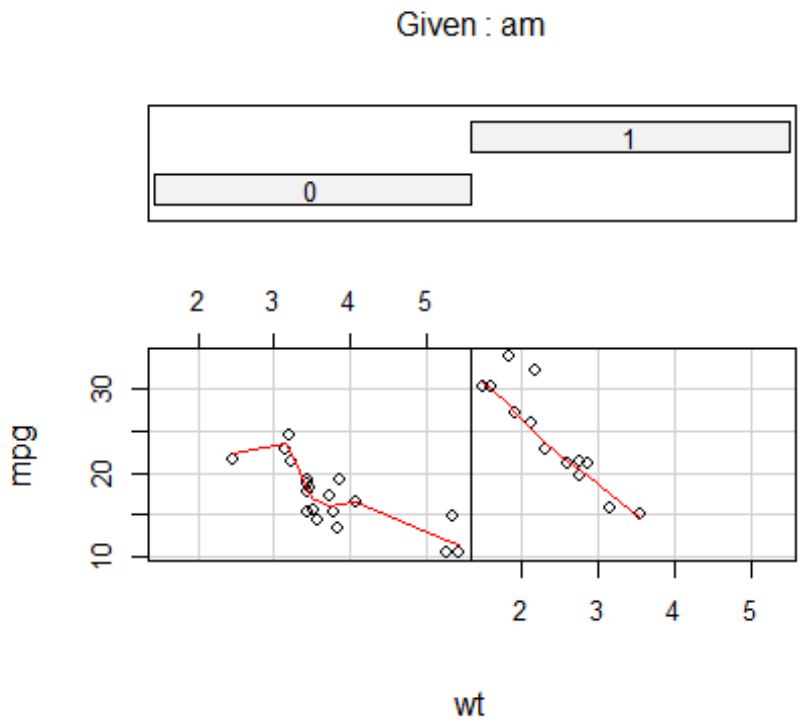
```
coplot(mpg ~ hp|am, data = mt, panel = panel.smooth, rows = 1)
```



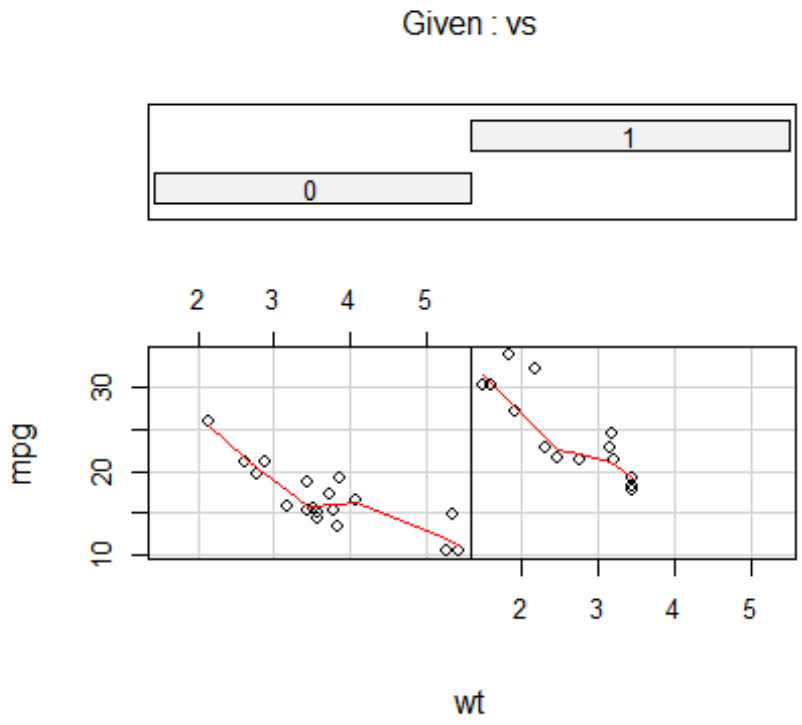

```
coplot(mpg ~ hp|vs, data = mt, panel = panel.smooth, rows = 1)
```




```
coplot(mpg ~ wt|am, data = mt, panel = panel.smooth, rows = 1)
```




```
coplot(mpg ~ wt|vs, data = mt, panel = panel.smooth, rows = 1)
```



The exploratory plots show that "vs" and "hp" are interacted, "am" and "wt" as well. So we need to consider the possible interaction between variables. Since we has only 32 observations, it's easy to be overfitting if we include too much interactions in the model. So Let' only include 6 two-way interactions.

```
model8 <- lm(mpg~hp+wt+am+vs+hp:wt+hp:am+hp:vs+wt:am+wt:vs+am:vs,data=mt)
summary(model8)
```

```
##
## Call:
## lm(formula = mpg ~ hp + wt + am + vs + hp:wt + hp:am + hp:vs +
##     wt:am + wt:vs + am:vs, data = mt)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -2.80  -1.25  -0.30   1.26   4.09
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.86380   19.01139    1.83    0.081 .
## hp          -0.05714    0.09118   -0.63    0.538
## wt          -3.96010    5.01896   -0.79    0.439
## am1          9.92703    9.91709    1.00    0.328
## vs1          3.41649   12.84581    0.27    0.793
## hp:wt        0.00933    0.02342    0.40    0.694
## hp:am1       0.02072    0.02989    0.69    0.496
## hp:vs1      -0.05685    0.04070   -1.40    0.177
```

```
## wt:am1      -4.30303    3.03391   -1.42    0.171
## wt:vs1       1.15551    3.24960    0.36    0.726
## am1:vs1      0.54155    4.28745    0.13    0.901
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.25 on 21 degrees of freedom
## Multiple R-squared:  0.906,  Adjusted R-squared:  0.861
## F-statistic: 20.2 on 10 and 21 DF,  p-value: 1.37e-08
```

The result is bad so let's just delete interactions with highest P-values and keep doing this to the new models until at least "hp", "wt" and "am" all have lower than 10% P-values. First, delete "am:vs".

```
model8.1 <- update(model8,.~.-am:vs)
model8.2 <- update(model8.1,.~.-wt:vs)
model8.3 <- update(model8.2,.~.-hp:wt)
model8.4 <- update(model8.3,.~.-hp:am)
model8.5 <- update(model8.4,.~.-hp)
```

Interesting thing happens as model8.4 give a good credibility and explanability except for "hp". Does this mean "hp" is unneccessary? If get rid of it, there must be another variable to represent the information that "hp" is giving and only "wt" is capable. The correlation between "hp" and "wt" is

```
cor(mt$hp,mt$wt)
```

```
## [1] 0.6587
```

A high correlation, so I would like to try get rid of "hp".

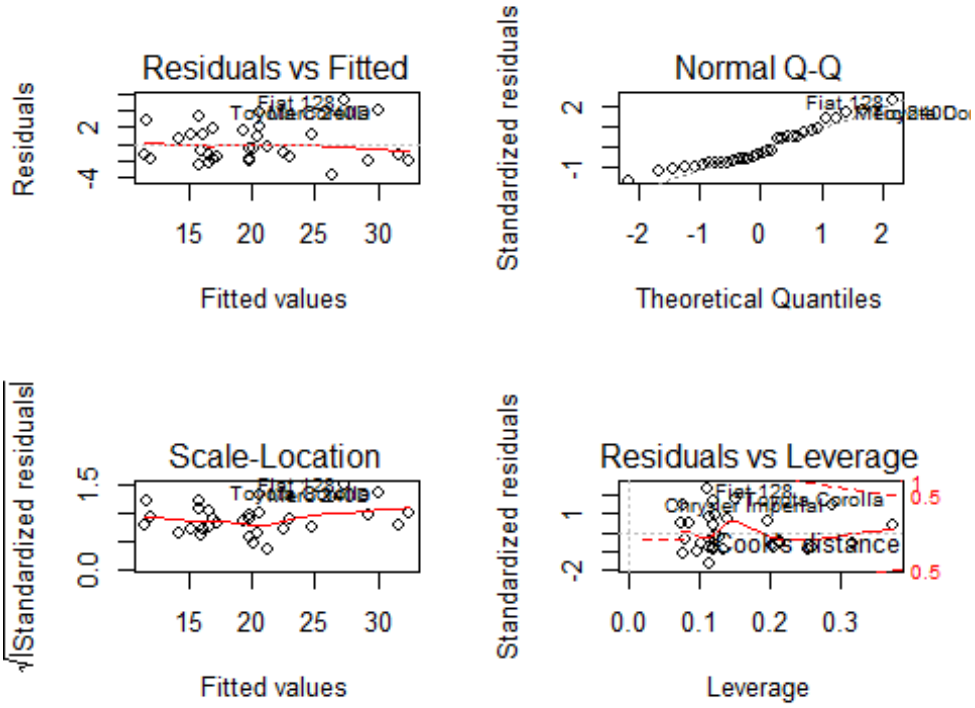
```
model9 <- lm(mpg~wt+vs+am+wt:am,data=mt)
summary(model9)
```

```
##
## Call:
## lm(formula = mpg ~ wt + vs + am + wt:am, data = mt)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.615 -1.682 -0.762  1.293  5.101
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   26.254      3.346    7.85  1.9e-08 ***
## wt            -2.703      0.818   -3.30  0.00270 **
## vs1            2.930      1.095    2.68  0.01249 *
## am1           14.320      3.866    3.70  0.00096 ***
## wt:am1        -4.663      1.329   -3.51  0.00160 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2.35 on 27 degrees of freedom
## Multiple R-squared:  0.868,  Adjusted R-squared:  0.849
## F-statistic: 44.4 on 4 and 27 DF,  p-value: 1.7e-11
```

Looks perfect, now the predictors are creditable and adjusted R2 are high. This one has potential to be the chosen one. Next step is to look at the linearity and normality.

```
par(mfrow=c(2,2))
plot(model9)
```



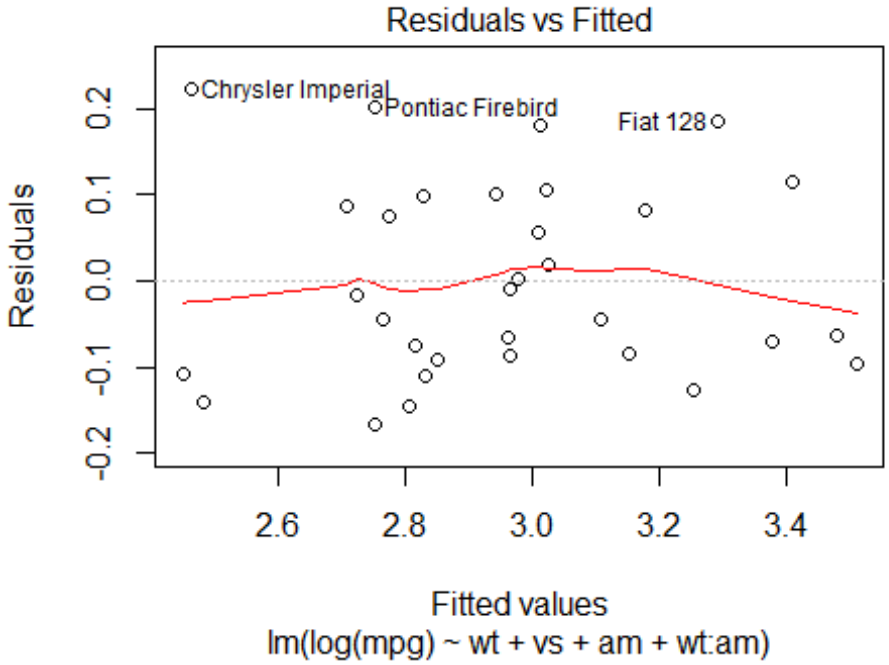
The plots show two problems: 1. QQ plot show that the model isn't normal. 2. There are outliers. For the normality problem, we can try using log(mpg) as the response; for outliers, we need to look at data.

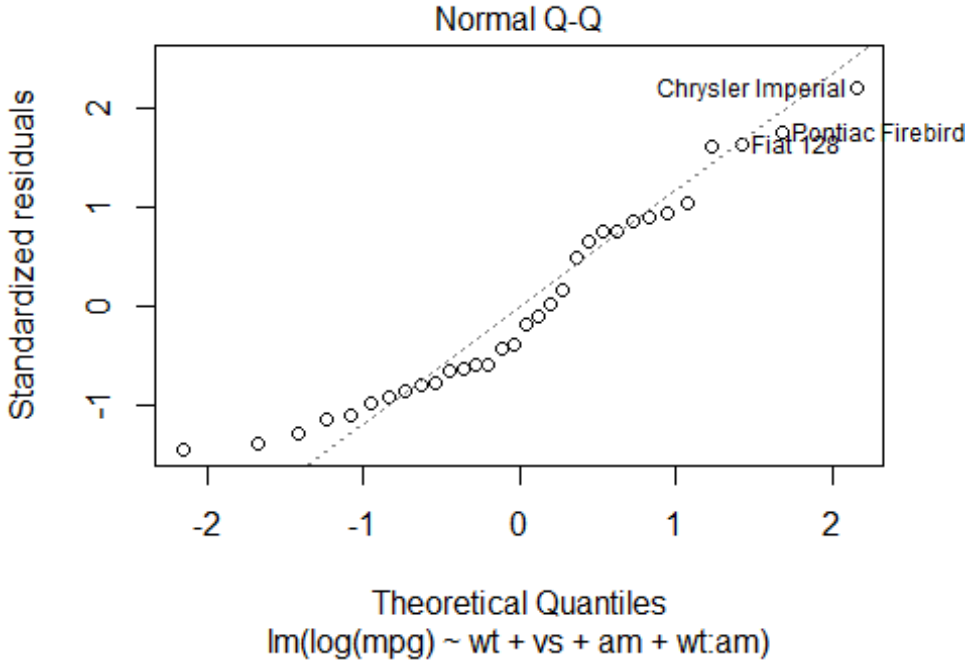
1. try to cure normality

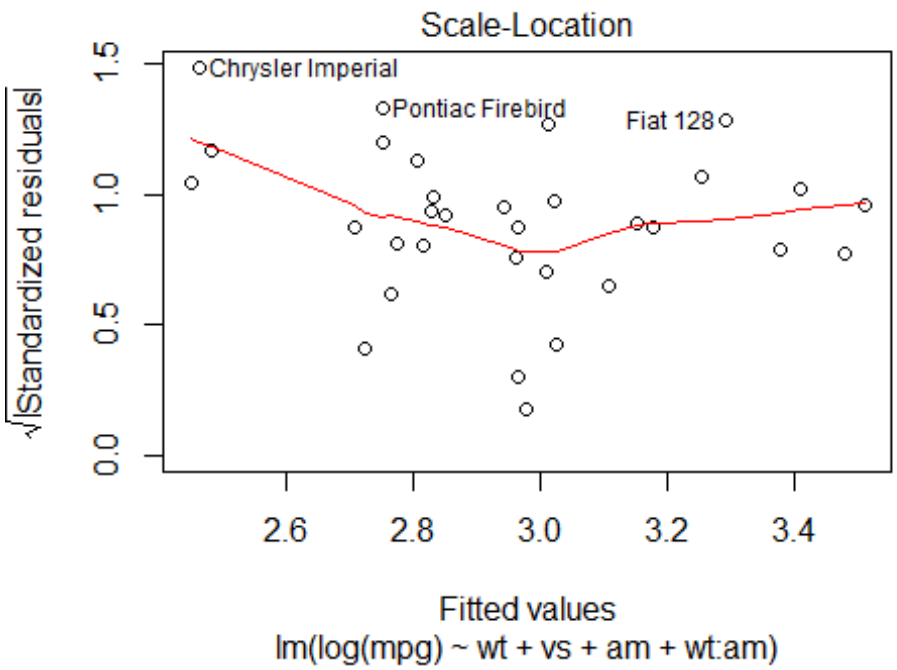
```
model9.log <- lm(log(mpg)~wt+vs+am+wt:am,data=mt)
summary(model9.log)

##
## Call:
## lm(formula = log(mpg) ~ wt + vs + am + wt:am, data = mt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.1656 -0.0869 -0.0298  0.0906  0.2228
##
```

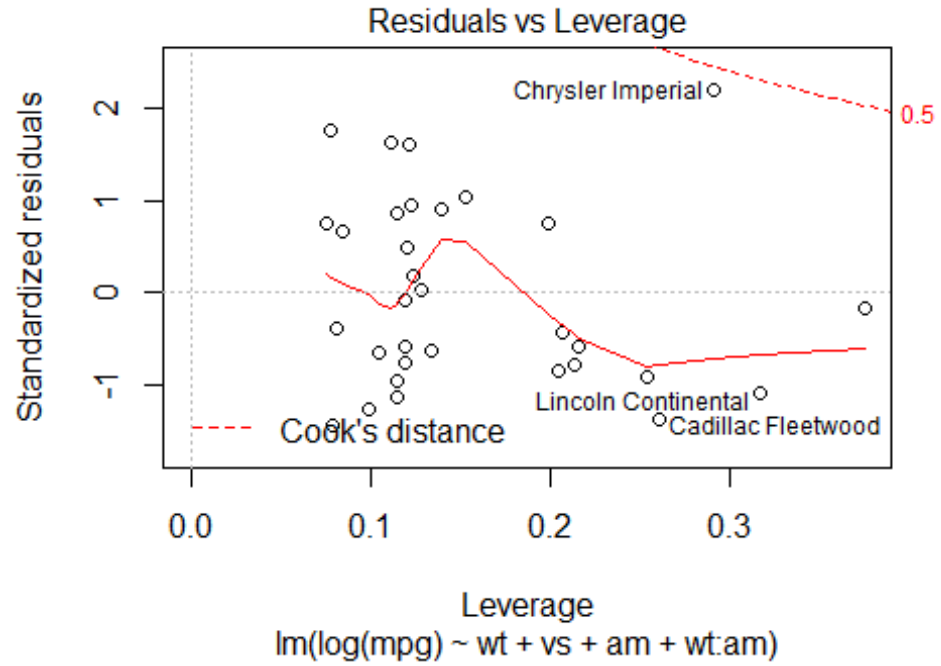
```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.4891     0.1712   20.37  < 2e-16 ***
## wt           -0.1916     0.0419   -4.57  9.6e-05 ***
## vs1           0.1349     0.0560    2.41    0.023 *
## am1           0.3641     0.1979    1.84    0.077 .
## wt:am1       -0.1247     0.0680   -1.83    0.078 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.12 on 27 degrees of freedom
## Multiple R-squared:  0.858,  Adjusted R-squared:  0.837
## F-statistic: 40.9 on 4 and 27 DF,  p-value: 4.36e-11
```

```
plot(model9.log)
```


Residuals vs Fitted
Chrysler Imperial
Pontiac Firebird
Fiat 128
Residuals
0.2
0.1
0.0
-0.1
-0.2
2.6
2.8
3.0
3.2
3.4
Fitted values
lm(log(mpg) ~ wt + vs + am + wt:am)


Normal Q-Q
Chrysler Imperial
Pontiac Firebird
Fiat 128
Standardized residuals
2
1
0
-1
-2
-1
0
1
2
Theoretical Quantiles
lm(log(mpg) ~ wt + vs + am + wt:am)



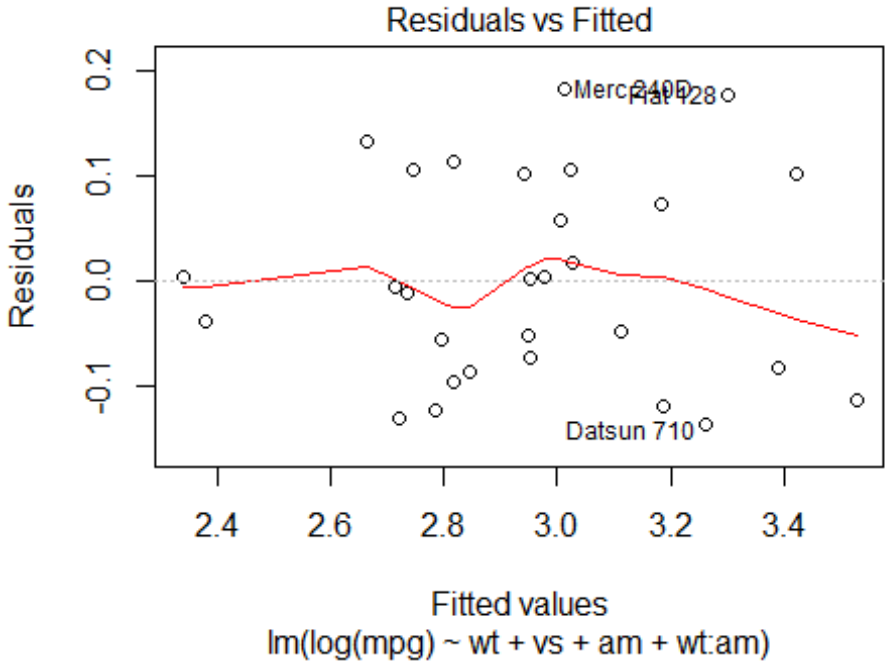


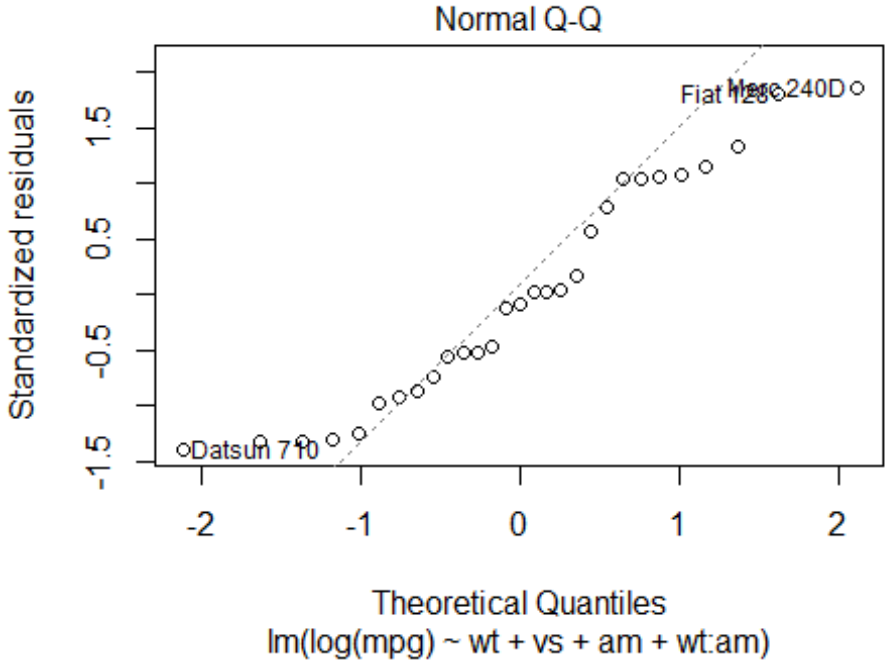
Normality is better. Ok let's not be greedy and turn to outliers.

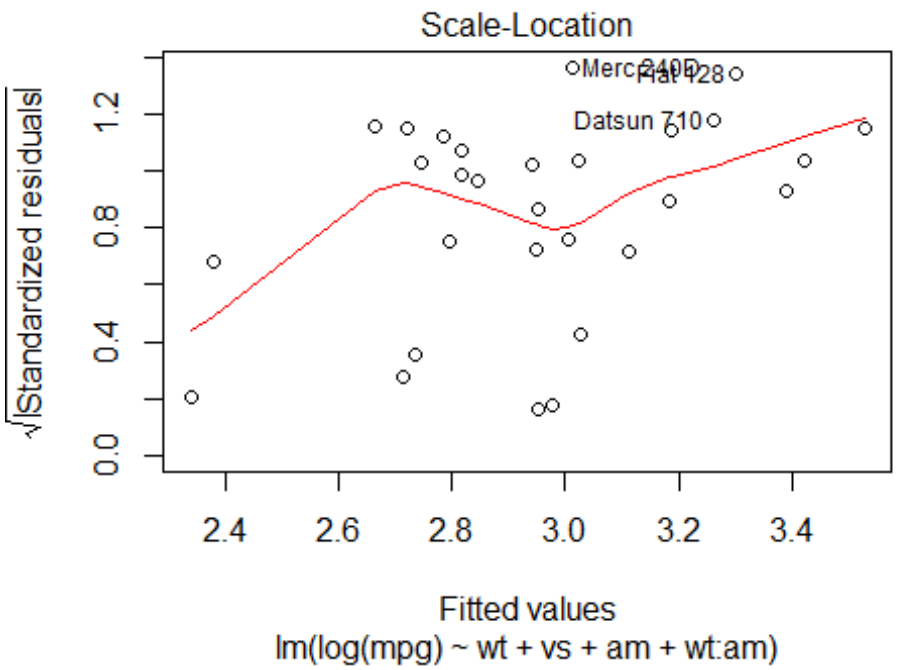
There are 5 car models are classified as outlier but three of them are shown in 3 plots and one is shown only in one plot. Since we only have 32 observations, the less outlier we taken

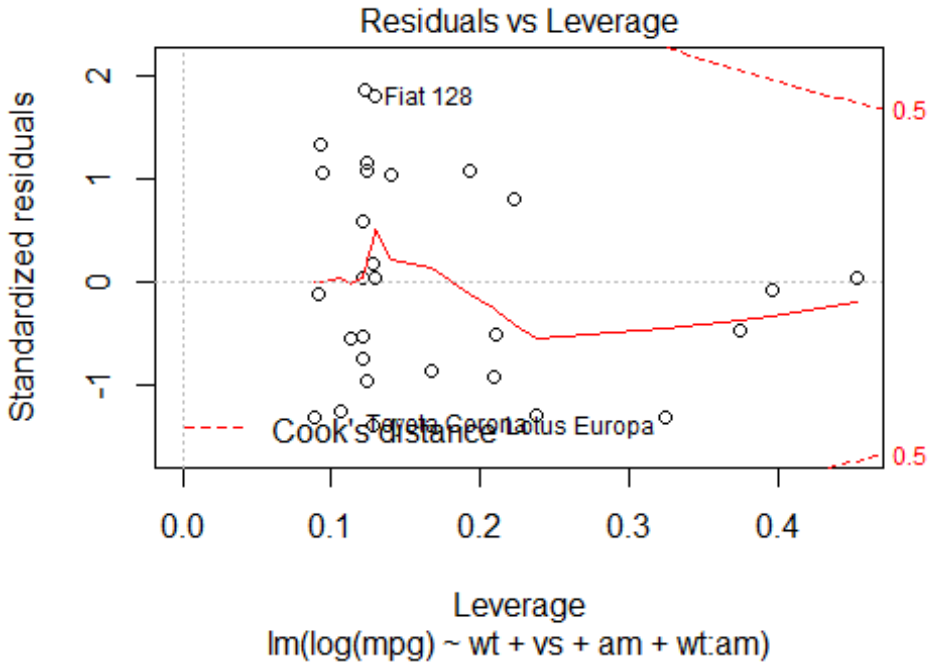
out the better, I will only take out 3 of them to see their impact on model9.log. 3 outliers are:Chrysler Imperial, No.17; Pontiac Firebird, No.25; and Fiat 128, No.19.

```
model9.log.o <- lm(log(mpg)~wt+vs+am+wt:am,data=mt[-c(17,19,25),])
plot(model9.log.o)
```


Residuals vs Fitted
Residuals
Merc 240D
Fiat 128
Datsun 710
0.2
0.1
0.0
-0.1
2.4
2.6
2.8
3.0
3.2
3.4
Fitted values
lm(log(mpg) ~ wt + vs + am + wt:am)


Normal Q-Q
Standardized residuals
Fiat 128
Merc 240D
Datsun 710
1.5
0.5
-0.5
-1.5
-2
-1
0
1
2
Theoretical Quantiles
lm(log(mpg) ~ wt + vs + am + wt:am)





The normality get worse and new outliers come out. Perhaps, we can't take the outliers out as the car markets are just natural to have outliers! Let's look at the so-called outliers:

```
outliers <- mt[c(17,19,25),c("am","vs","mpg","wt")]
outliers
```

```
##                   am vs  mpg    wt
## Chrysler Imperial  0  0 14.7 5.345
## Honda Civic        1  1 30.4 1.615
## Pontiac Firebird   0  0 19.2 3.845
```

```
require(plyr)
```

```
## Loading required package: plyr
```

```
means <-ddply(mt, .(am,vs),summarize, MeanMpg = round(mean(mpg),3),MeanWt = r
ound(mean(wt),3))
means
```

```
##   am vs MeanMpg MeanWt
## 1  0  0   15.05  4.104
## 2  0  1   20.74  3.194
## 3  1  0   19.75  2.857
## 4  1  1   28.37  2.028
```

The comparison between outliers and the cars with same am and vs, in terms of mean mpg and wt, show that Chrysler Imperial become outlier because it's too heavy, Honda Civic too light and Pontiac Firebird too fuel-efficient. But those qualities are just normal in car market and doesn't support any special treatment in data analytics like this. Therefore, model9.log won't take out outliers and this makes it the chosen model I present to you.

Hope you like it! Notice I am not even a driver, I am pretty happy to finally finish this assignment. If you are a car lover, please tell me does this model make sense and any thoughts on choosing a second-hand car because I am going to America and gonna need a car!