Course 7 - Regression Models Project

Grejell Segura

8/21/2017

Excutive Summary

This report investigated the difference of manual and automatic transmission cars by looking at their miles/galloon measures. Other variables are also investigated to find possible interactions and confounding effects. The data used is the mtcars taken from the package "datasets".

Load Libraries

I used a number of libraries to help me with the visualizations and testing.

```r
set.seed(4324)
library(ggplot2)
library(datasets)
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 3.3.3
```

```r
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 3.3.3
```

```r
library(car)
```

```
## Warning: replacing previous import 'lme4::sigma' by 'stats::sigma' when
## loading 'pbkrtest'
```

```r
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 3.3.3
```

```r
library(caret)
```

Exploratory Data Analysis

First, let us check the data structure.

```r
str(mtcars)
```

```
## '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 ...
```

The data has 32 observations with 11 variables.

Now let us convert some of the variables to factors.

```
mtcars[, c("cyl", "vs", "am", "gear", "carb")] <- lapply(mtcars[, c("cyl", "vs", "am", "gear", "carb")]
```

Then let us examine the data by looking at the summary of all the variables.

```
summary(mtcars)
```

```
##       mpg        cyl         disp              hp              drat      
##  Min.   :10.40   4:11   Min.   : 71.1   Min.   : 52.0   Min.   :2.760  
##  1st Qu.:15.43   6: 7   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080  
##  Median :19.20   8:14   Median :196.3   Median :123.0   Median :3.695  
##  Mean   :20.09          Mean   :230.7   Mean   :146.7   Mean   :3.597  
##  3rd Qu.:22.80          3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920  
##  Max.   :33.90          Max.   :472.0   Max.   :335.0   Max.   :4.930  
##        wt             qsec       vs     am     gear   carb  
##  Min.   :1.513   Min.   :14.50   0:18   0:19   3:15   1: 7  
##  1st Qu.:2.581   1st Qu.:16.89   1:14   1:13   4:12   2:10  
##  Median :3.325   Median :17.71                 5: 5   3: 3  
##  Mean   :3.217   Mean   :17.85                        4:10  
##  3rd Qu.:3.610   3rd Qu.:18.90                        6: 1  
##  Max.   :5.424   Max.   :22.90                        8: 1  
```

To have an initial insight, it is best to visualize the data. We will focus first on mgp and am variables as this is our main concern on this problem. The result can be in Figure 1 below.

```
plot.1 <- ggplot(mtcars, aes(x = mpg, fill = am)) + geom_histogram(bins = 32) + labs(title = "Figure 1"
plot.1
```

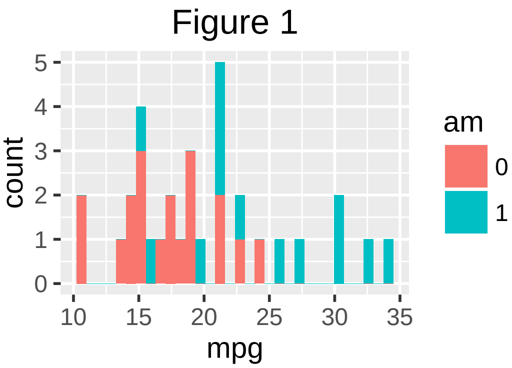
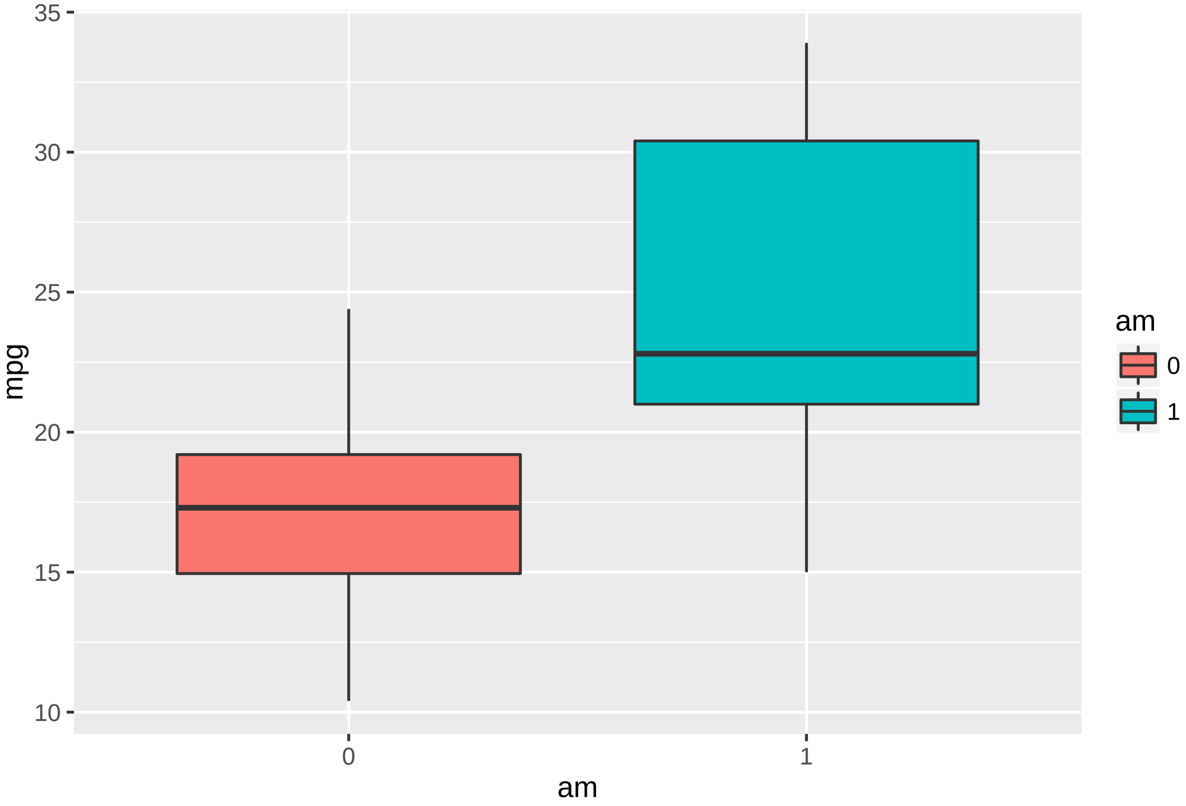


Figure 1 shows the distribution of mpg. The colors indicate the 2 types of transmission - manual is shown red while automatic is shown in green. We can see that there is some sort of separation between 2 transmission types where automatic transmission mostly lies to the right and manual lies to the left of the graph. A boxplot was also created to see if there is visual difference between the groups. Looking at Figure 2, we can see a difference of mpg between the 2 types of transmissions. Automatic is visually higher here.

```
plot.2 <- ggplot(mtcars, aes(x = am, y = mpg, fill = am)) + geom_boxplot() + labs(title = "Figure 2")
plot.2
```

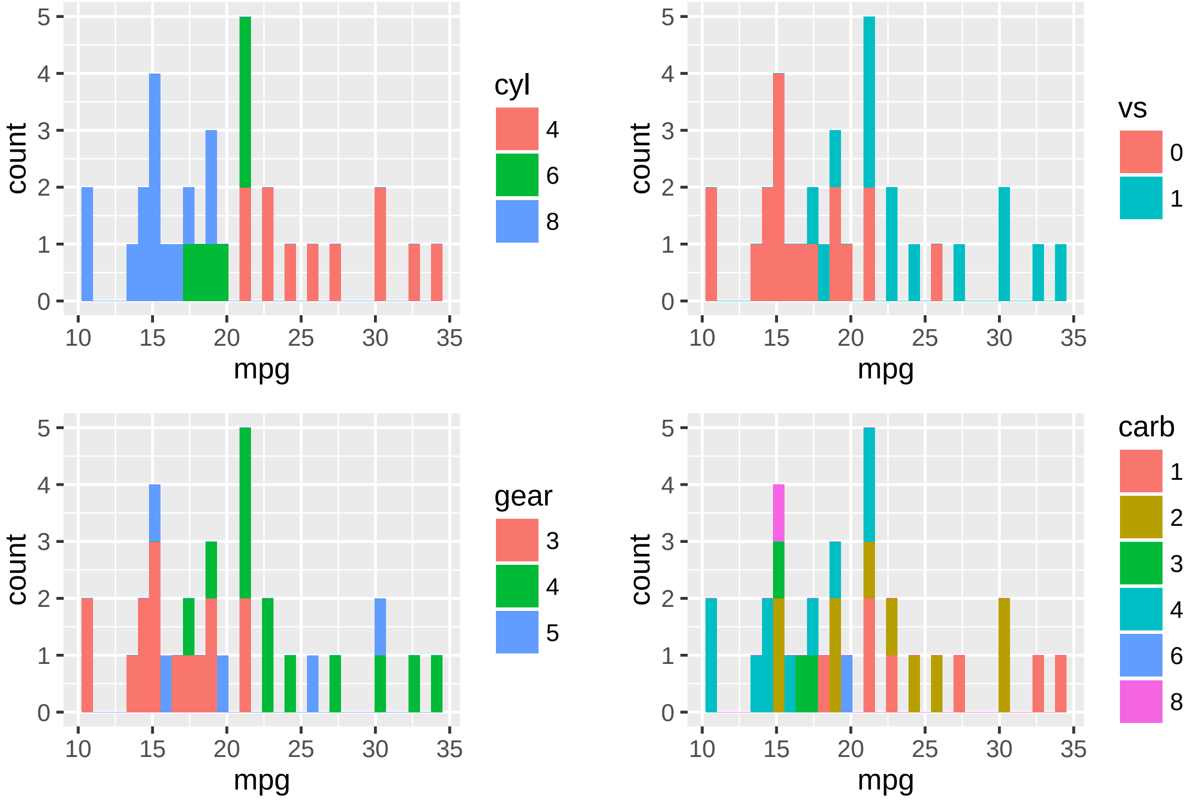
Figure 2



Another chart was created to examine the relationship of mpg to other categorical variables. The result is showed below as Figure 3.

```
plot.cyl <- ggplot(mtcars, aes(x = mpg, fill = cyl)) + geom_histogram(bins = 32)
plot.vs <- ggplot(mtcars, aes(x = mpg, fill = vs)) + geom_histogram(bins = 32)
plot.gear <- ggplot(mtcars, aes(x = mpg, fill = gear)) + geom_histogram(bins = 32)
plot.carb <- ggplot(mtcars, aes(x = mpg, fill = carb)) + geom_histogram(bins = 32)
grid.arrange(plot.cyl, plot.vs, plot.gear, plot.carb, ncol = 2, top = "Figure 3")
```

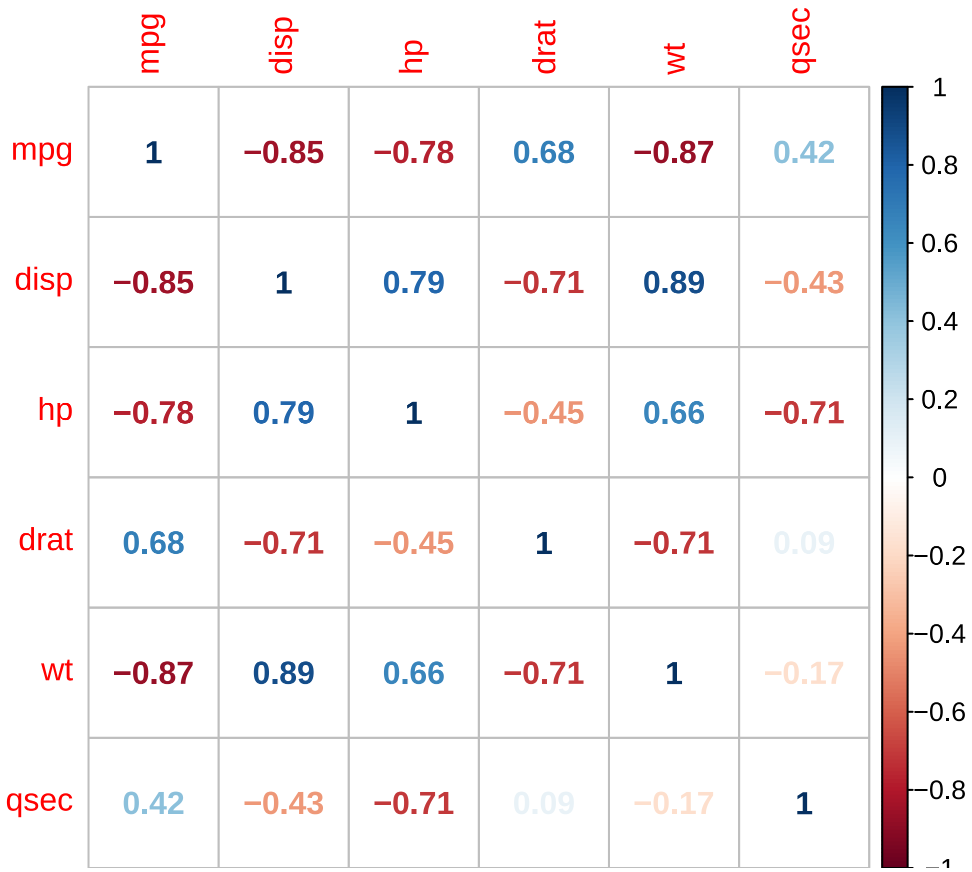
Figure 3



Here we can see that there is also some sort of separation between mpg vs. cyl, mpg vs. vs, and mpg vs. gear. As for mpg vs. carb, there is no clear separation to it.

We are done visualizing and examining mpg against the categorical variables. Next is we try to see if it also has some sort of relation with respect to the numerical variables. Let us see by looking at the correlation graph below.

```r
cor <- cor(mtcars[, c("mpg", "disp", "hp", "drat", "wt", "qsec")])
corrplot(cor, method = "number")
```



	mpg	disp	hp	drat	wt	qsec
mpg	1	−0.85	−0.78	0.68	−0.87	0.42
disp	−0.85	1	0.79	−0.71	0.89	−0.43
hp	−0.78	0.79	1	−0.45	0.66	−0.71
drat	0.68	−0.71	−0.45	1	−0.71	0.09
wt	−0.87	0.89	0.66	−0.71	1	−0.17
qsec	0.42	−0.43	−0.71	0.09	−0.17	1

Looking at the heatmap, mpg has high negative correlation with disp, hp, and wt. On the other hand, drat has relatively high positive correlation with mpg, qsec meanwhile has the lowest correlation with mpg.

Inferential Analysis

Let us now examine the relationships by having an inferential analysis. The main objective is to see if there is an evidence of the difference of 2 types of transmission. We also want to quantify this effect by fitting a model.

We will use a t-test to determine if there is an evidence of difference in the mpg means for the 2 different types. Looking back at the graph, we have an assumption that automatic transmission has more mpg mean so we will have the hypothesis testing in this manner.
Ho : there is no difference of mpg means between 2 types of transmissions Ha : automatic transmission has greater mean than manual transmission

```
model.1 <- t.test(mpg ~ am, mtcars, alternative = "greater")
model.1
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.9993
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -10.57662       Inf
## sample estimates:
```

```
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

The t-test result shows p-value = 0.00069 which very small. Hence we reject the null hypothesis and declare that there is a statistical difference between the 2 transmissions and that automatic transmission has greater mpg mean than its manual counterpart. Ofcourse we dont want to stop the investigation at this point as we also wish to quantify the proven difference. We proceed by linear modelling.

```
model.1 <- lm(mpg ~ am, mtcars)
summary(model.1)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am1            7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

The model shows the significance of the coefficient in am1. This means there is a 7.245 mpg difference between automatic transmission and manual transmission. However, we should also consider that there are other confounding variables that may also affect the this effect. Therefore, we will proceed on creating a model consisting of some of the variables that may also affect mpg.
We wish to build the best model by starting with our model.1 and nesting it by adding 1 variable at a time.

```
model.2 <- update(model.1, mpg ~ am + cyl, mtcars)
model.3 <- update(model.1, mpg ~ am + cyl + hp, mtcars)
model.4 <- update(model.1, mpg ~ am + cyl + hp + wt, mtcars)
anova(model.1, model.2, model.3, model.4)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl
## Model 3: mpg ~ am + cyl + hp
## Model 4: mpg ~ am + cyl + hp + wt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     28 264.50  2    456.40 39.286 1.388e-08 ***
## 3     27 197.20  1     67.30 11.585  0.002164 **
## 4     26 151.03  1     46.17  7.949  0.009081 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

I stopped adding variables at model.4 as we don't find any significant difference by adding the remaining variables. Let us examine model.4 then.

```
summary(model.4)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + hp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## am1          1.80921    1.39630   1.296  0.20646
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

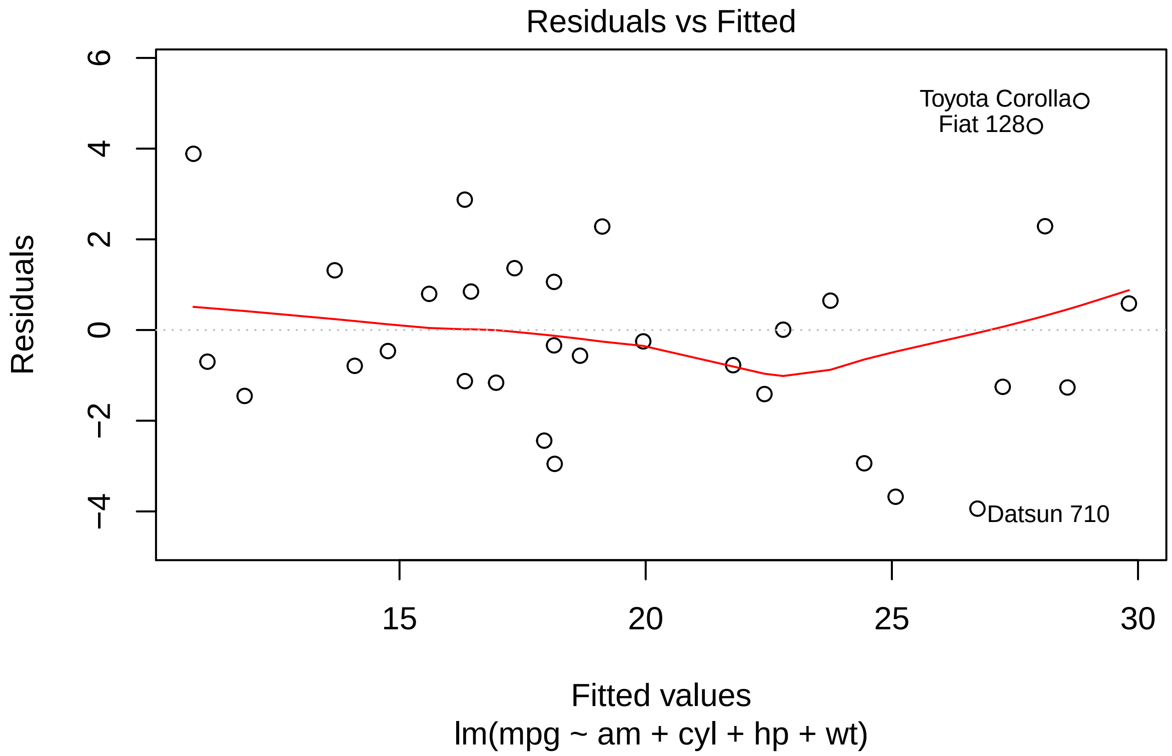
```
AIC(model.1, model.2, model.3, model.4)
```

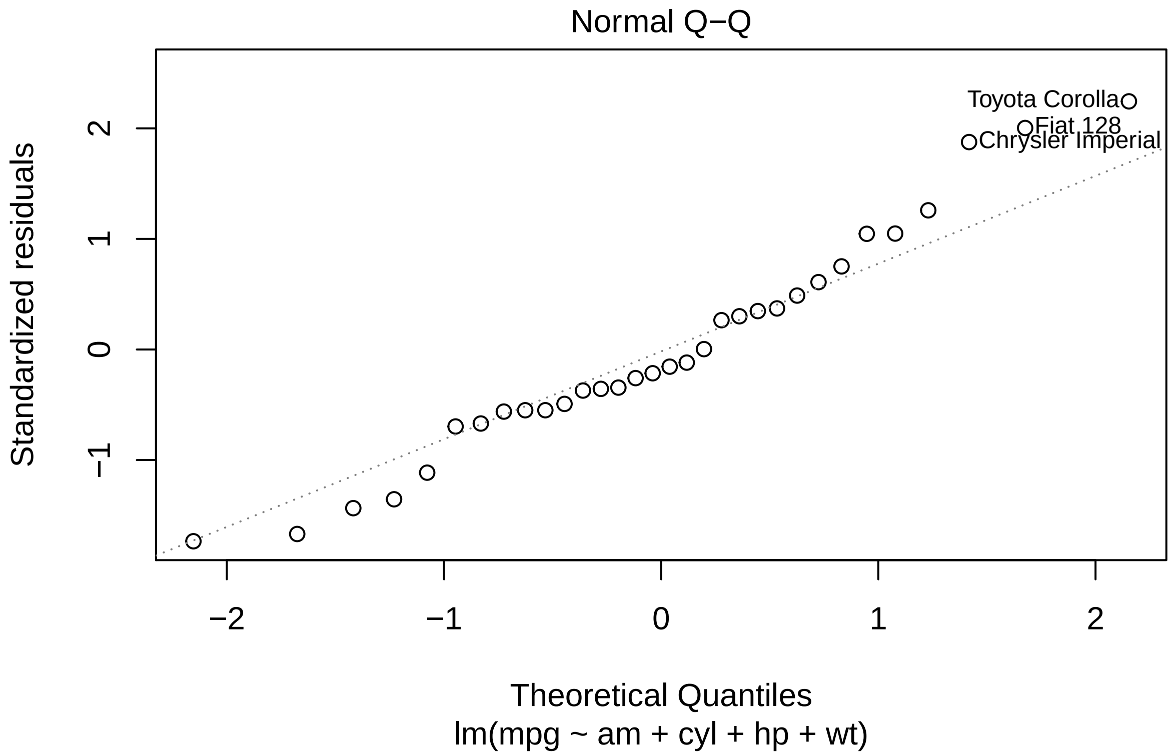
```
##         df      AIC
## model.1  3 196.4844
## model.2  5 168.3989
## model.3  6 161.0033
## model.4  7 154.4669
```

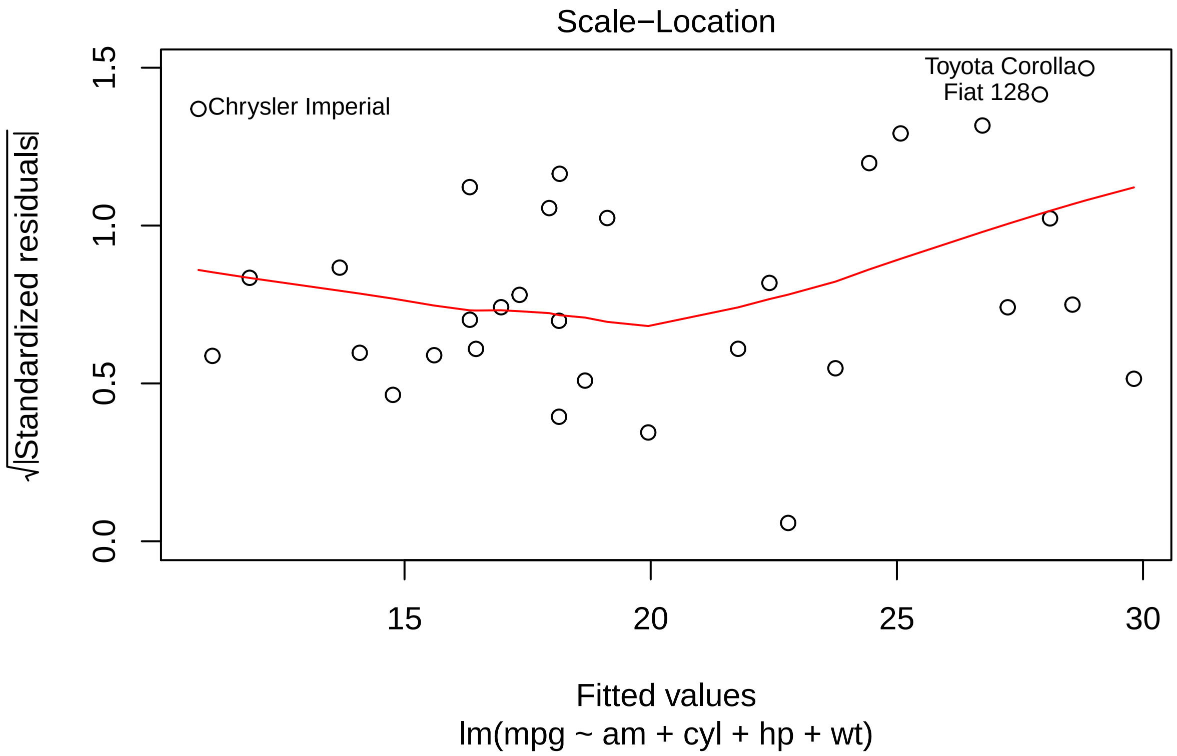
The summary shows that the model explains 84.01% of the variance of mpg. The model also has the lowest Akaike Information Criterion (AIC) which means it is the best prediction model among the 4 considered.

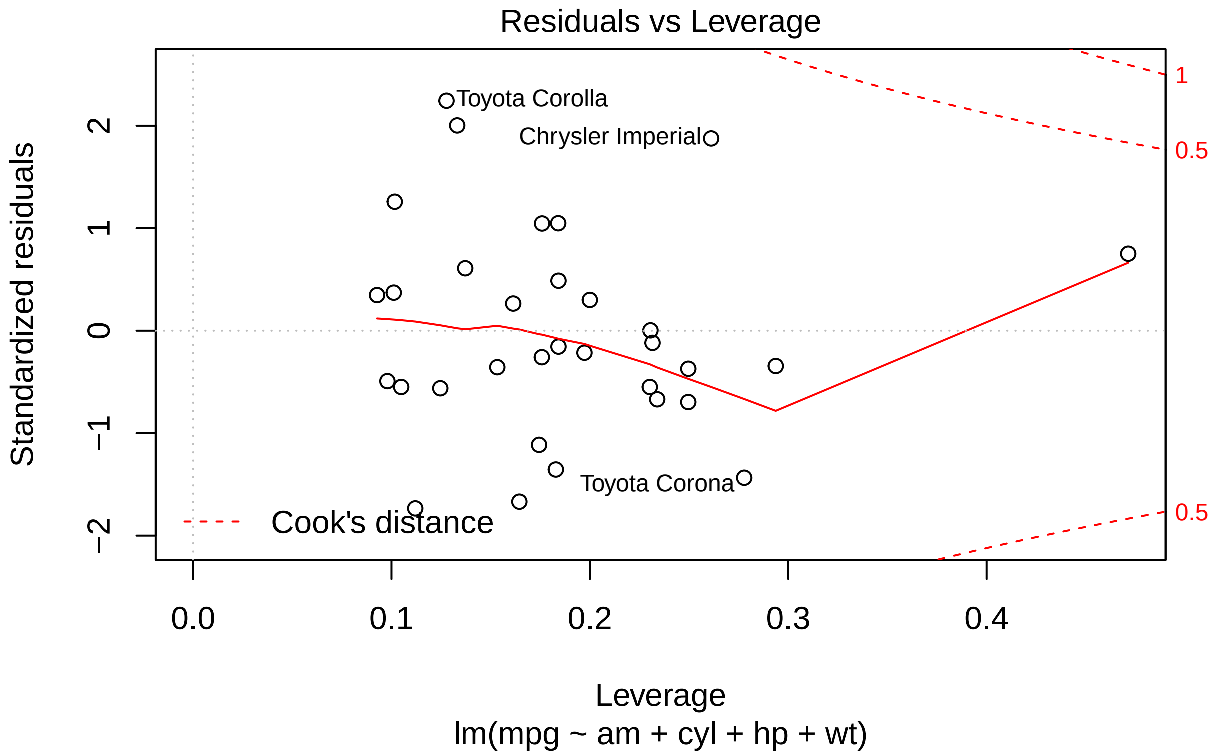
The following are the residual plots to investigate the goodness of fit of model.4.

```
plot(model.4)
```


Residuals vs Fitted
Residuals
6
4
2
0
−2
−4
15
20
25
30
Fitted values
lm(mpg ~ am + cyl + hp + wt)
Toyota Corolla
Fiat 128
Datsun 710


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(mpg ~ am + cyl + hp + wt)
−2
−1
0
1
2
−1
0
1
2
Toyota Corolla
Fiat 128
Chrysler Imperial


Scale−Location
1.5
1.0
0.5
0.0
√|Standardized residuals|
Chrysler Imperial
Toyota Corolla
Fiat 128
15
20
25
30
Fitted values
lm(mpg ~ am + cyl + hp + wt)



```
grid.arrange
```

```
## function (..., newpage = TRUE)
## {
##     if (newpage)
##         grid.newpage()
##     g <- arrangeGrob(...)
##     grid.draw(g)
##     invisible(g)
## }
## <environment: namespace:gridExtra>
```

The plots shows that the residuals are randomly scattered and distributed. This is a good indication.

To make sure, the following diagnostic tests are also considered to see if the model is good enough.

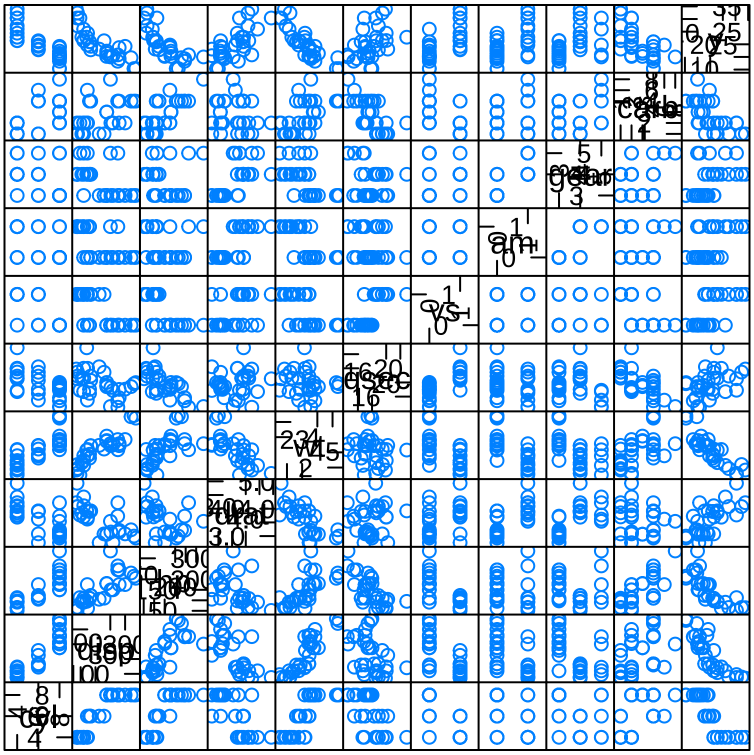
```
dwtest(model.4)
```

```
##
##  Durbin-Watson test
##
## data:  model.4
## DW = 1.8088, p-value = 0.1567
## alternative hypothesis: true autocorrelation is greater than 0
```

The dwtest is the Durbin-Watson test which tests the existence of autocorrelation in the residuals. As observed, the p-value is greater than 0.5 which means there is no evidence that an autocorrelation exist.

Appendix

```r
featurePlot(mtcars[, -1], mtcars$mpg, plot = "pairs")
```



Scatter Plot Matrix