Regression Models Course Project

Tan Woei Ming

15 April 2016

Executive Summary

mtcars is the data extracted from the 1974 Motor Trend US magazine, consist of fuel consumption (mpg) and 10 other design and performance parameters (cyl,disp,hp,drat,wt,qsec,vs,am,gear and carb) [type ?mtcars for more information]

Data exploration is performed using basic plot to find the relationship between the parameters and mpg. Using the properties of linear regression, we are able to find any significant change of parameters by keeping others constant, when performing full parameters model fit.

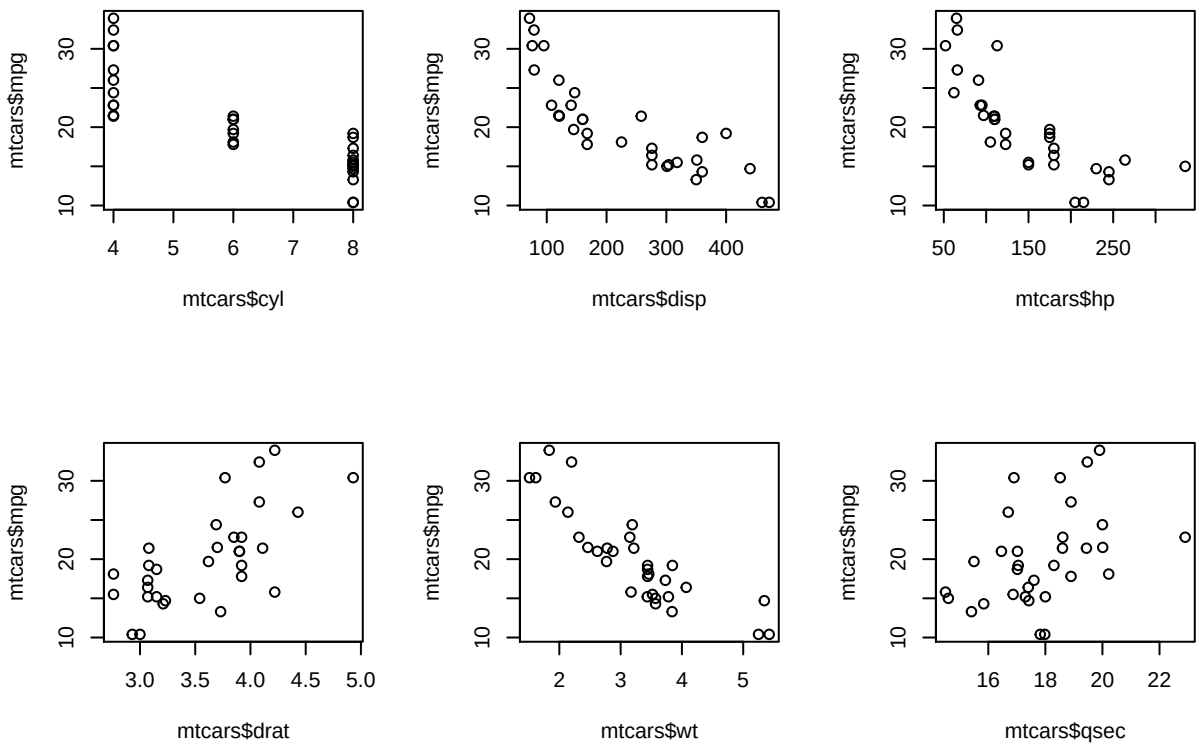
Instead of over-fit (full parameters), **we narrow down to wt, qsec and am, using the p-value in the full parameters model fit**. We fit mpg using those parameters and also find the **significant different (95% confidence level) between the transmission type (am), where manual transmission is 2.94 miles/gallon (keeping others no change) higher than the automatic one**.

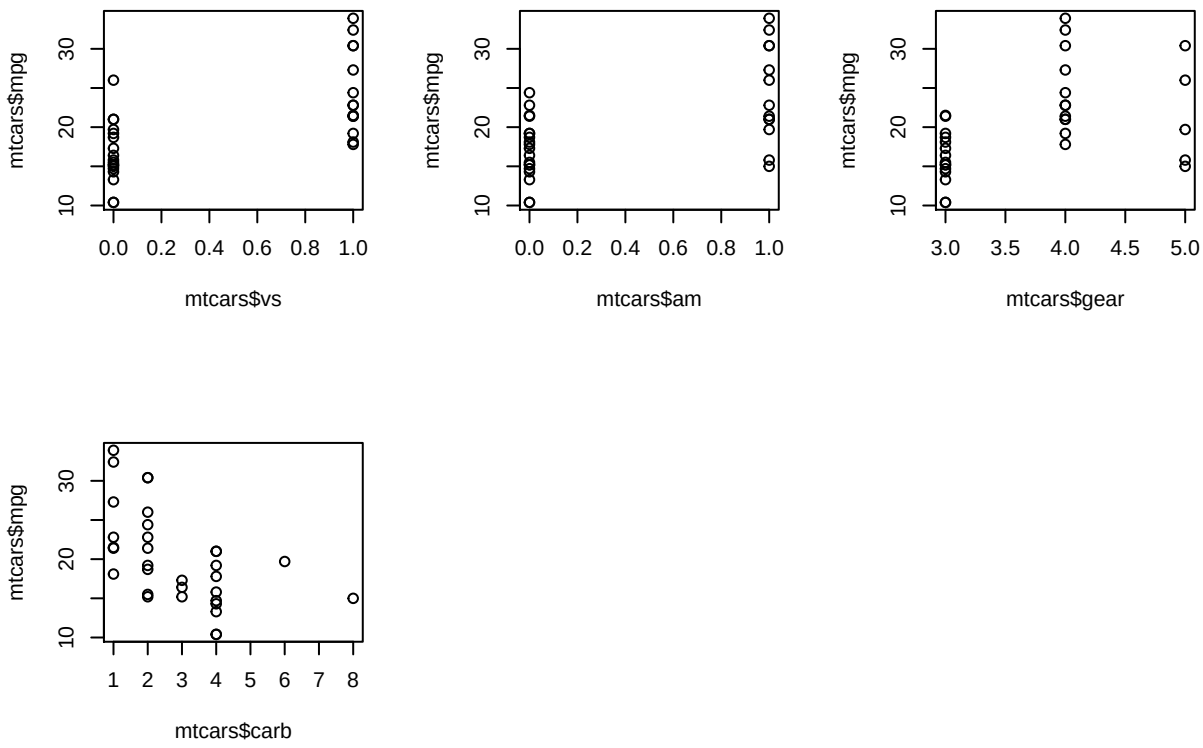
A. Data Exploratory and Basic Data Analysis

First, we want to understand more on each parameters and the data type. Below shows that:

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

We also make plot between each parameters and mpg to briefly understand the relationship. From the plot, we find that there are multiple parameters have the relationship with mpg.


mtcars$mpg
mtcars$cyl
mtcars$disp
mtcars$hp
mtcars$drat
mtcars$wt
mtcars$qsec



B. Model Selection and Fitting

We could start by over-fit the model and find out the residual range, so that we could do an iterative model fit by adding the relavent parameters one by one and keeping the residual in range.

First, we try to fit all parameters, the **residual range (-3.450644,4.627094) and Adjusted R-squared 0.8066**. This is the residual range and R-squared target we want to achieve later when we select the parameters for the model.

```
fit_full<-lm(mpg~.,data=mtcars)
#model fit summary
summary(fit_full)
```

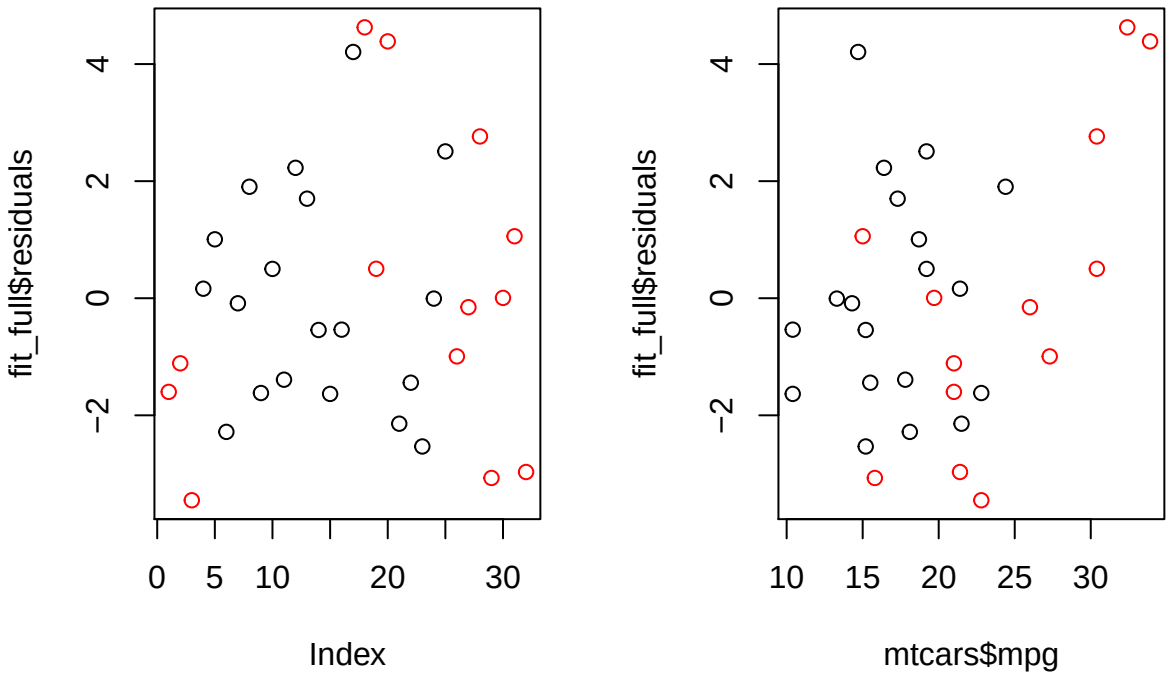
```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
```

```
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

```r
#residual range
range(fit_full$residuals)
```

```
## [1] -3.450644  4.627094
```

```r
#residual plot
par(mfrow=c(1,2))
plot(fit_full$residuals,col=factor(mtcars$am))
plot(fit_full$residuals~mtcars$mpg,col=factor(mtcars$am))
```



Second, we also try to fit just the transmission type (am). Obviously the residual is higher than the full parameters, in the **range (-9.392308,9.507692) and Adjusted R-squared is low at 0.3385**. If we look at the residual by mpg plot, observed there is a consistent different between the automatic and manual transmission, meaning we need to add more parameters to the model to bring down the residual.

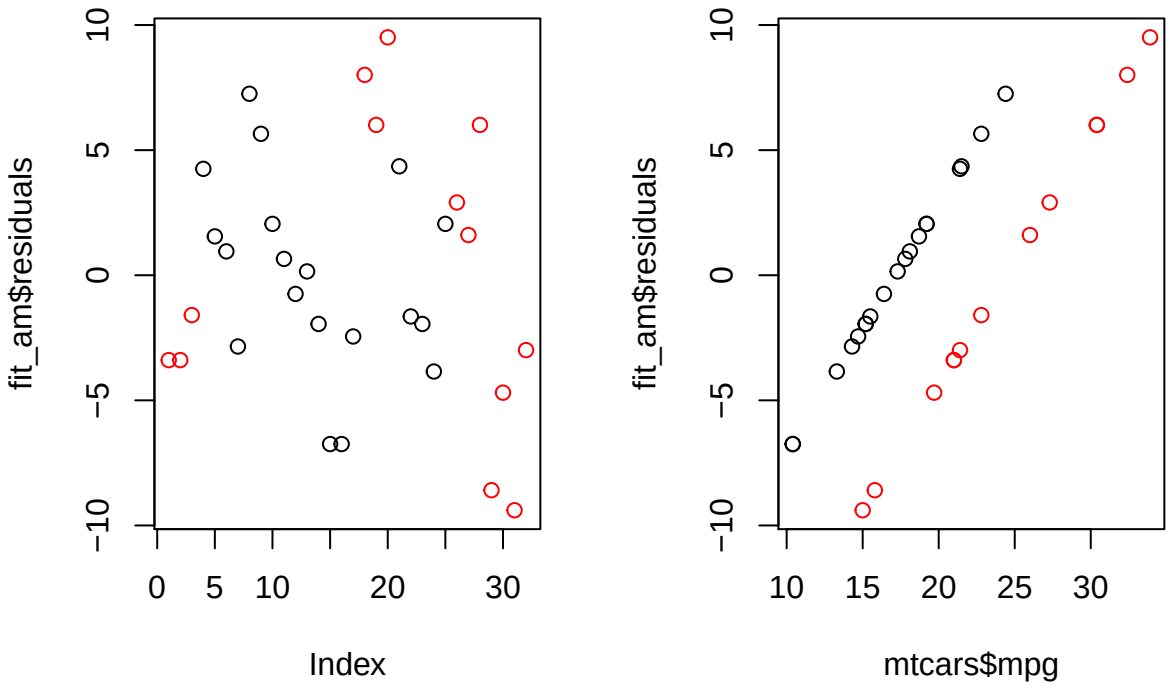
```
fit_am<-lm(mpg~factor(am),data=mtcars)
#model fit summary
summary(fit_am)
```

```
##
## Call:
## lm(formula = mpg ~ factor(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 ***
## factor(am)1    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
```

```
#residual range
range(fit_am$residuals)
```

```
## [1] -9.392308  9.507692
```

```
#residual plot
par(mfrow=c(1,2))
plot(fit_am$residuals,col=factor(mtcars$am))
plot(fit_am$residuals~mtcars$mpg,col=factor(mtcars$am))
```



In order to add more relavent paremeters, we take the top 3 parameters in the full parameter fit based on the p-value: wt, qsec and am. Fit those parameters to the model and find the **residual range in (-3.481067,4.660998) and Adjusted R-squared 0.8336**, which is close to our target residual range and R-squared (in the over-fit model). **In other words, we have achieved the balance between fitting too little parameters and over-fit situation. So, we narrow down the parameters from 10 to just 3, with the acceptable residual range and R-squared.**

```
fit_opt<-lm(mpg~wt+qsec+factor(am),data=mtcars)
#model fit summary
summary(fit_opt)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + factor(am), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## factor(am)1   2.9358     1.4109   2.081 0.046716 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

```
#residual range
range(fit_opt$residuals)
```

```
## [1] -3.481067  4.660998
```

C. Coefficient Interpretation

Review the last model fit (fit_opt), we find that wt, qsec and am fit pretty well with the acceptable residual range and R-squared. Now lets look at the coefficient:

To find the slope of automatic transmission (am=0), is the summation of Intercept, wt and qsec, which is 6.927163. The slope of manual transmission (am=1), is the summation of Intercept, wt,qsec and am, which is 9.863. The p-value between automatic and manual transmission (am=0 and am=1) is 0.047.

The test show that there is a different in transmission with 95% confidence level. The different is 2.935837.

```
#coefficient and p-value
summary(fit_opt)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## factor(am)1  2.935837  1.4109045  2.080819 4.671551e-02
```

```
#slope of am=0
sum(summary(fit_opt)$coef[1:3,1])
```

```
## [1] 6.927163
```

```
#slope of am=1
sum(summary(fit_opt)$coef[1:4,1])
```

```
## [1] 9.863
```

```
#the p-value between am=0 and am=1
summary(fit_opt)$coef[4,4]
```

```
## [1] 0.04671551
```

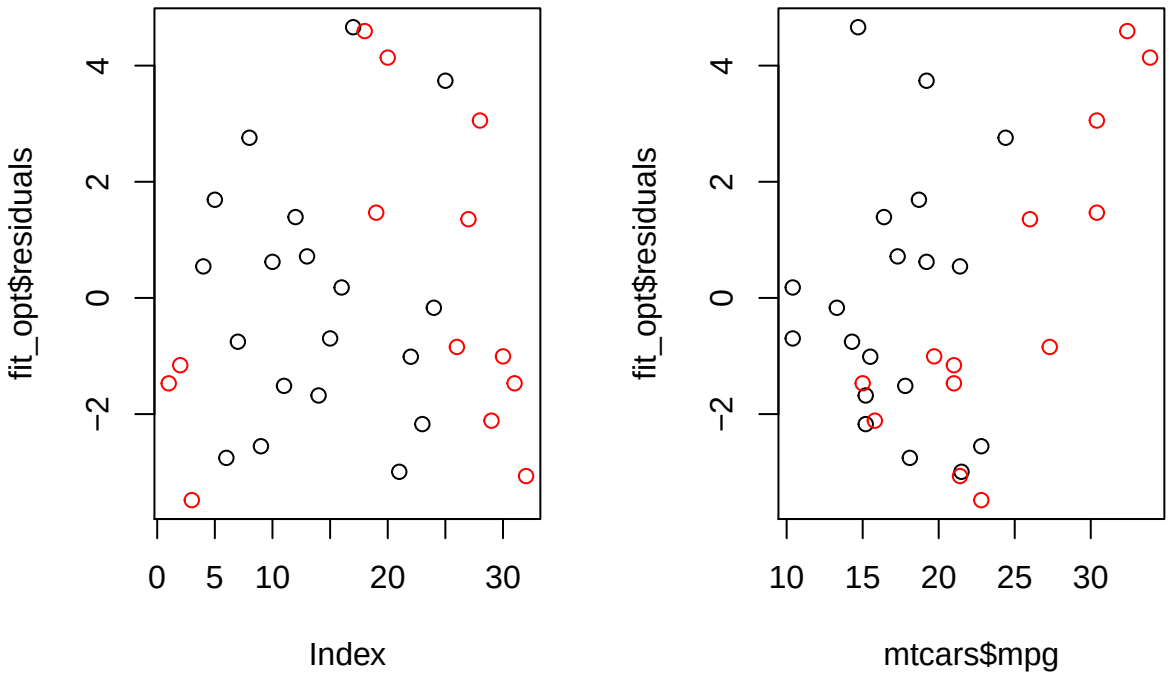
```
#the different in mpg between am=0 and am=1
sum(summary(fit_opt)$coef[1:4,1])-sum(summary(fit_opt)$coef[1:3,1])
```

```
## [1] 2.935837
```

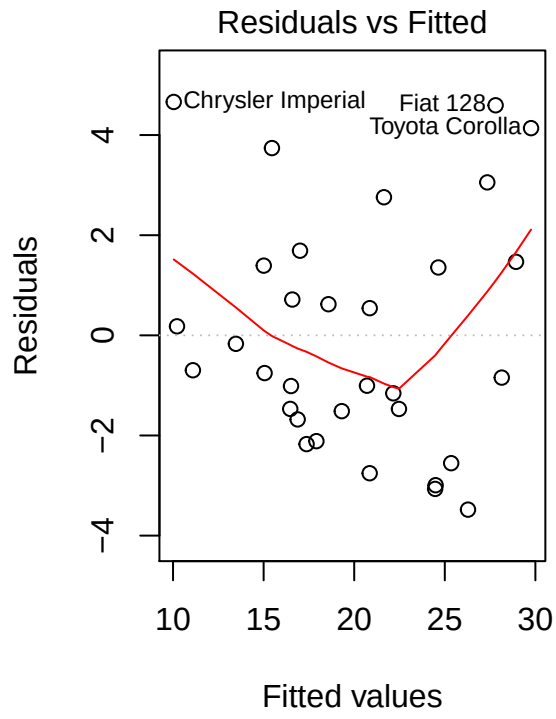
D. Residual Analysis

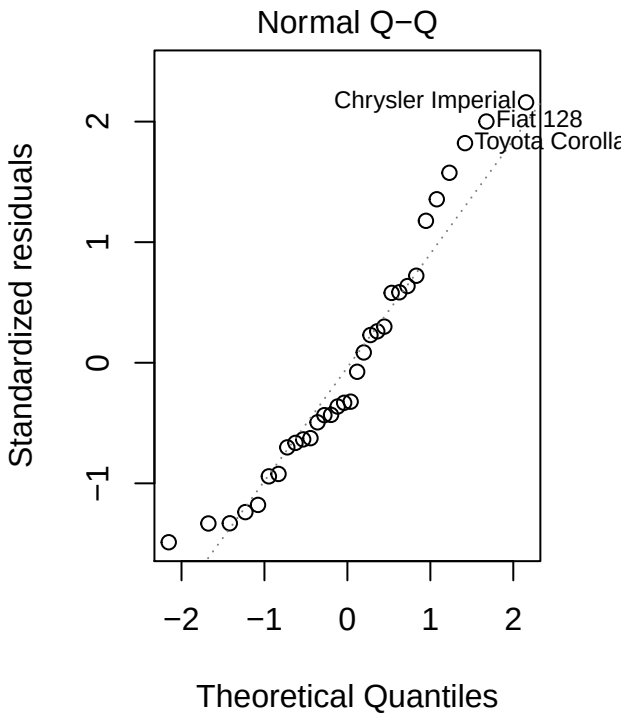
The residual of the fit_opt model is random across the mpg. There is no pattern observed in the residual plot.

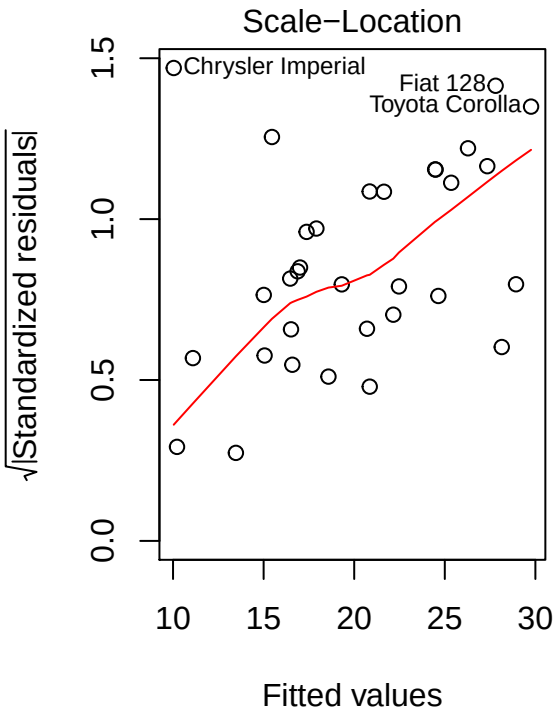
```
#residual plot
par(mfrow=c(1,2))
plot(fit_opt$residuals,col=factor(mtcars$am))
plot(fit_opt$residuals~mtcars$mpg,col=factor(mtcars$am))
```

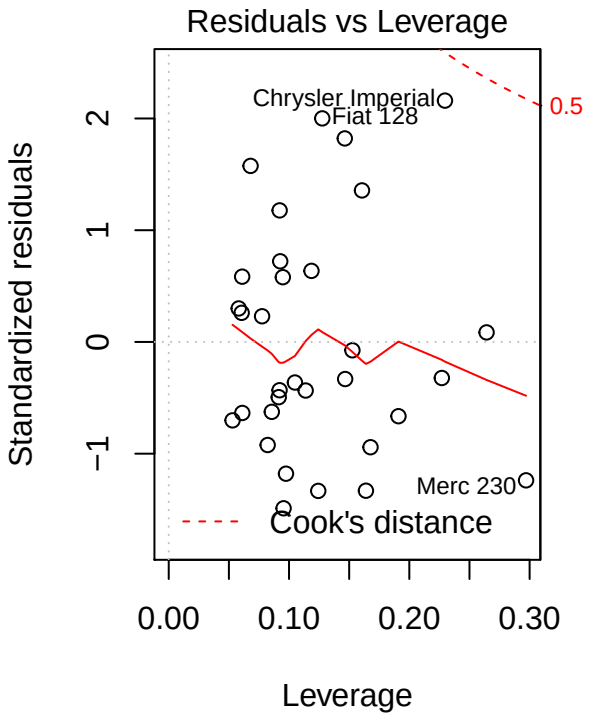



```
plot(fit_opt)
```


Residuals vs Fitted
Chrysler Imperial
Fiat 128
Toyota Corolla
Residuals
Fitted values


Normal Q−Q
Chrysler Imperial
Fiat 128
Toyota Corolla
Standardized residuals
Theoretical Quantiles


Scale−Location
Chrysler Imperial
Fiat 128
Toyota Corolla
√|Standardized residuals|
Fitted values


Residuals vs Leverage
Chrysler Imperial
Fiat 128
0.5
Merc 230
Cook's distance
Standardized residuals
Leverage

DataScience - Rgression Models http://github.com/twming/datascience-regressionmodels