Regression Models Course Project

Yulong Wang

19/09/2021

```
setwd("/Users/yulong/GitHub/Statistical-Inference-Course-Project")
```

Instructions

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

- "Is an automatic or manual transmission better for MPG"
- "Quantify the MPG difference between automatic and manual transmissions"

Overview

To answer these questions, we conducted exploratory data analysis (EDA) and used hypothesis testing and linear regression. We have established simple and multiple linear regression analysis. However, the results of the multivariate regression model are more promising because it includes the potential effects of other variables on MPG.

Using the model selection strategy, we found that among all variables, weight and quarter-mile time (acceleration) have a significant effect on quantifying the difference in MPG between automatic and manual transmission cars.

Data Processing

```
library(datasets)
data(mtcars)
```

For the purpose of this analysis, we use the mtcars data set, which is a data set extracted from the American Automobile Trends magazine in 1974. It contains the fuel consumption of 32 cars (1973-74 models) and 10 of the car design and performance. Aspects. The following is a brief description of the variables in the data set:

It consists of 32 observations on 11 variables.

[, 1] mpg Miles/(US) gallon [, 2] cyl Number of cylinders [, 3] disp Displacement (cu.in.) [, 4] hp Gross horsepower [, 5] drat Rear axle ratio [, 6] wt Weight (lb/1000) [, 7] qsec 1/4 mile time [, 8] vs V/S [, 9] am Transmission (0 = automatic, 1 = manual) [,10] gear Number of forward gears [,11] carb Number of carburetors

```
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 ...
```

All the data is the num format, also, show the fisrt several rows of the data below:

```
library(knitr)
library(printr)
```

```
## Registered S3 method overwritten by 'printr':
##   method                from
##   knit_print.data.frame rmarkdown
```

```
kable(head(mtcars),align = 'c')
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

Please note that each row of mtcars represents a car model, which we can see in the row name. Each column is an attribute of the car, such as the number of miles per gallon (or fuel efficiency), the number of cylinders, the displacement (or volume) of the car's engine (in cubic inches), whether the car is an automatic transmission or a manual transmission, etc.

Exploratory data analyses

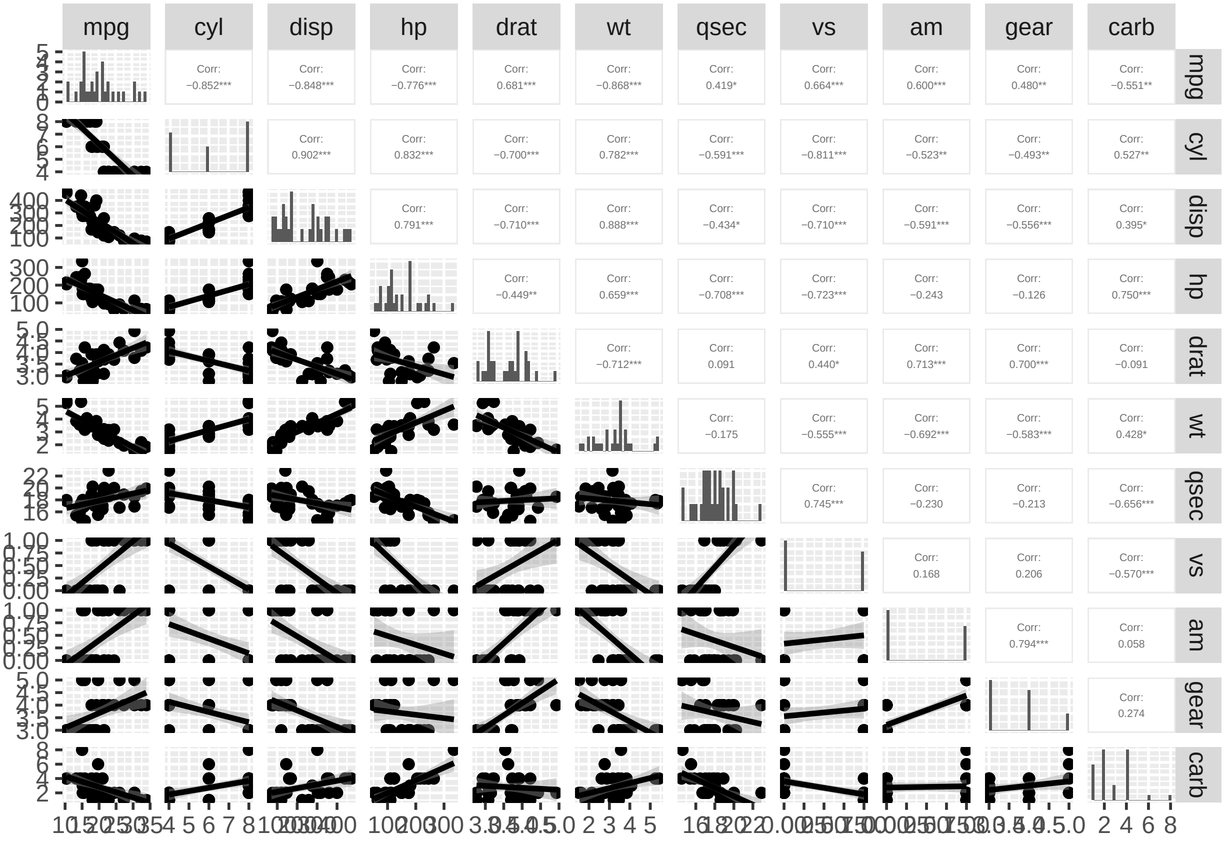
```
library(GGally)
```

```
## Loading required package: ggplot2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

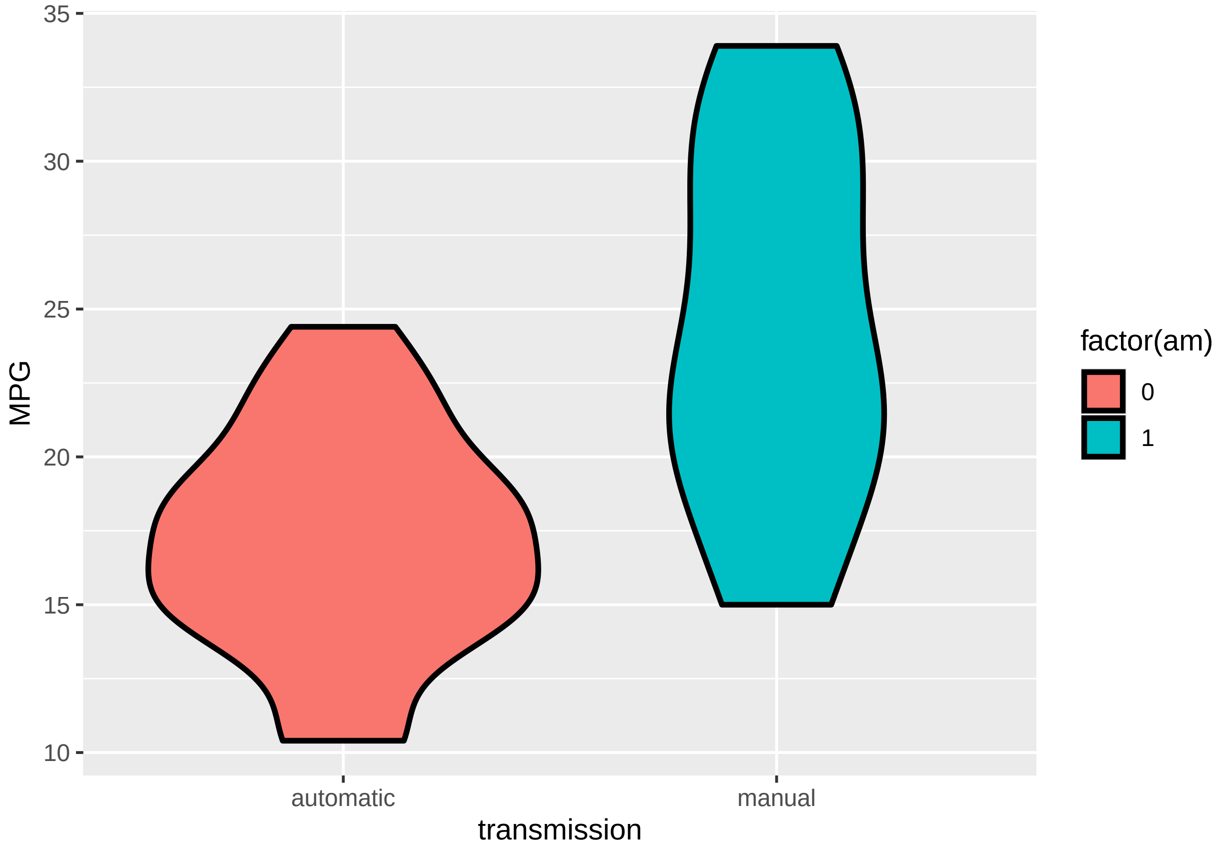
```
library(ggplot2)
ggpairs(mtcars,
        upper = list(continuous = wrap("cor", size = 1.5)),
        lower = list(continuous = "smooth"),
        diag = list(continuous = "bar"),
        axisLabels='show')
```

```
## Warning in check_and_set_ggpairs_defaults("diag", diag, continuous =
## "densityDiag", : Changing diag$continuous from 'bar' to 'barDiag'

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



It is also worthwhile check how MPG varies by automatic versus manual transmission. For that purpose we create a Violin plot of MPG by automatic and manual transmissions. In our dataset 0 represents an automatic transmission and 1 means a manual transmission.

```
library(stats)
ggplot(mtcars, aes(y=mpg, x=factor(am, labels = c("automatic", "manual")), fill=factor(am)))+
        geom_violin(colour="black", size=1)+
        xlab("transmission") + ylab("MPG")
```



We can form a clear hypothesis from this visualization: Compared with manual cars, autonomous cars seem to have lower miles per gallon and therefore lower fuel efficiency. But this obvious pattern may happen randomly—that is, we just happened to choose a group of inefficient automatic cars and a group of more efficient manual cars. Therefore, to check whether this is the case, we must use a statistical test.

Model fitting and hypothesis testing

Two samples t-test

We are interested to know if an automatic or manual transmission better for MPG. So first we test the hypothesis that cars with an automatic transmission use more fuel than cars with a manual transmission. To compare two samples to see if they have different means, we use two sample T-test.

```
test <- t.test(mpg ~ am, data= mtcars, var.equal = FALSE,
               paired=FALSE ,conf.level = .95)
result <- data.frame( "t-statistic"  = test$statistic,
                       "df" = test$parameter,
                        "p-value"  = test$p.value,
                        "lower CL" = test$conf.int[1],
                        "upper CL" = test$conf.int[2],
                        "automatic mean" = test$estimate[1],
                        "manual mean" = test$estimate[2],
                        row.names = "")
kable(x = round(result,3),align = 'c')
```

t.statistic	df	p.value	lower.CL	upper.CL	automatic.mean	manual.mean
-3.767	18.332	0.001	-11.28	-3.21	17.147	24.392

p-value that shows the probability that this apparent difference between the two groups could appear by chance is very low. The confidence interval also describes how much lower the miles per gallon is in manual cars than it is in automatic cars. We can be confident that the true difference is between 3.2 and 11.3.

Simple linear regression model

We can also fit factor variables as regressors and come up with thing like analysis of variance as a special case of linear regression models. From the "dummy variables" point of view, there's nothing special about analysis of variance (ANOVA). It's just linear regression in the special case that all predictor variables are categorical. Our factor variable in this case is Transmission (am).

```
mtcars$amfactor <- factor(mtcars$am, labels = c("automatic", "manual"))
summary(lm(mpg ~ factor(amfactor), data = mtcars))$coef
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.147368	1.124602	15.247492	0.000000
factor(amfactor)manual	7.244939	1.764422	4.106127	0.000285

All the estimates provided here are in comparison with automatic transmission. The intercept of 17.14 is simply the mean MPG of automatic transmission. The slope of 7.24 is the change in the mean between manual transmission and automatic transmission. You can verify that from the plot as well. The p-value of 0.000285 for the mean MPG difference between manual and automatic transmission is significant. Therefore we conclude that according to this model manual transmission if more fuel efficient.

Fitting multivariable linear regression model

Modeling based on only one predictor variable does not seem to be sufficient and good enough as we have other predictor variables that might affect MPG and therefore affect the difference in MPG by transmission. So the univariate model in this case is only part of the picture. Therefore in this part of the analysis we use multivariable linear regression to develop a model that includes the effect of other variables.

Model selection procedure We want to know what combination of predictors will best predict fuel efficiency. Which predictors increase our accuracy by a statistically significant amount? We might be able to guess at the some of the trends from the graph, but really we want to perform a statistical test to determine which predictors are significant, and to determine the ideal formula for prediction.

Including variables that we should't have increases actuall standard errors of the regression variables.Thus we don't want to idly throw variables into the model. To confirm this fact, you can see below that if we include all the variables, not of them will a significant predictor of MPG (judging by p-value at the 95% confidence level).

```r
summary(lm(mpg ~ cyl+disp+hp+drat+wt+qsec+factor(vs)+factor(am)+gear+carb, data = mtcars))$coef
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.3033742	18.7178844	0.6573058	0.5181244
cyl	-0.1114405	1.0450234	-0.1066392	0.9160874
disp	0.0133352	0.0178575	0.7467585	0.4634887
hp	-0.0214821	0.0217686	-0.9868407	0.3349553
drat	0.7871110	1.6353731	0.4813036	0.6352779
wt	-3.7153039	1.8944143	-1.9611887	0.0632522
qsec	0.8210407	0.7308448	1.1234133	0.2739413
factor(vs)1	0.3177628	2.1045086	0.1509915	0.8814235
factor(am)1	2.5202269	2.0566506	1.2254035	0.2339897
gear	0.6554130	1.4932600	0.4389142	0.6652064
carb	-0.1994193	0.8287525	-0.2406258	0.8121787

Detecting collinearity A major problem with multivariate regression is collinearity. If two or more predictor variables are highly correlated, and they are both entered into a regression model, it increases the true standard error and you get a very unstable estimates of the slope. We can assess the collinearity by variance inflation factor (VIF). Lets look at the variance inflation factors if we throw all the variables into the model.

```r
library(car)
```

```
## Loading required package: carData
```

```r
fitvif <- lm(mpg ~ cyl+disp+hp+drat+wt+qsec+factor(vs)+factor(am)+gear+carb, data = mtcars)
kable(vif(fitvif),align = 'c')
```

	x
cyl	15.373833
disp	21.620241
hp	9.832037
drat	3.374620
wt	15.164887
qsec	7.527958
factor(vs)	4.965873
factor(am)	4.648487
gear	5.357452
carb	7.908747

Values for the VIF that are greater than 10 are considered large. We should also pay attention to VIf values between 5 and 10. At these point we might consider leaving only one of these variables in the model.

Stepwise selection method Among available methods we decided to perform stepwise selection to help us select a subset of variables that best explain the MPG. Please note that we also treat the vc variable as a categorical variable.

```
library(MASS)
fit <- lm(mpg ~ cyl+disp+hp+drat+wt+qsec+factor(vs)+factor(am)+gear+carb, data = mtcars)
step <- stepAIC(fit, direction="both", trace=FALSE)
summary(step)$coeff
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.617781	6.9595930	1.381946	0.1779152
wt	-3.916504	0.7112016	-5.506882	0.0000070
qsec	1.225886	0.2886696	4.246676	0.0002162
factor(am)1	2.935837	1.4109045	2.080819	0.0467155

```
summary(step)$r.squared
```

```
## [1] 0.8496636
```

This shows that in addition to transmission, weight of the vehicle as well as acceleration speed have the highest relation to explaining the variation in mpg. The adjusted R^2 is 84% which means that the model explains 84% of the variation in mpg indicating it is a robust and highly predictive model.

Nested likelihood ratio test If the models of interest are nested and without lots of parameters differentiating them, it's fairly uncontroversial to use nested likelihood ratio tests. So in order to verify the result of the stepwise selection model, we also perform this procedure below.

```
fit1 <- lm(mpg ~ factor(am), data = mtcars)
fit2 <- lm(mpg ~ factor(am)+wt, data = mtcars)
fit3 <- lm(mpg ~ factor(am)+wt+qsec, data = mtcars)
fit4 <- lm(mpg ~ factor(am)+wt+qsec+hp, data = mtcars)
fit5 <- lm(mpg ~ factor(am)+wt+qsec+hp+drat, data = mtcars)
anova(fit1, fit2, fit3, fit4, fit5)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
30	720.8966	NA	NA	NA	NA
29	278.3197	1	442.576902	72.5359307	0.0000000
28	169.2859	1	109.033768	17.8700375	0.0002579
27	160.0665	1	9.219469	1.5110205	0.2299925
26	158.6386	1	1.427847	0.2340163	0.6326111

As you can see, the result is consistent with stepwise selection model and adding any more variable in addition to wt, am and qsec will dramatically increase the variation in the model, and the p-value immediately becomes insignificant.

Fitting the final model Now using the selected variables, we can fit the final model.

```
finalfit <- lm(mpg ~ wt+qsec+factor(am), data = mtcars)
summary(finalfit)$coef
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.617781	6.9595930	1.381946	0.1779152
wt	-3.916504	0.7112016	-5.506882	0.0000070
qsec	1.225886	0.2886696	4.246676	0.0002162
factor(am)1	2.935837	1.4109045	2.080819	0.0467155

You can observe that all the variables now are statistically significant. This model explains 84% of the variance in miles per gallon (mpg). Now when we read the coefficient for am, we say that, on average, manual transmission cars have 2.94 MPGs more than automatic transmission cars. However this effect was much higher than when we did not adjust for weight and qsec.

Regression diagnostics

In this section, we perform some diagnostics on the final regression model.

Detecting collinearity

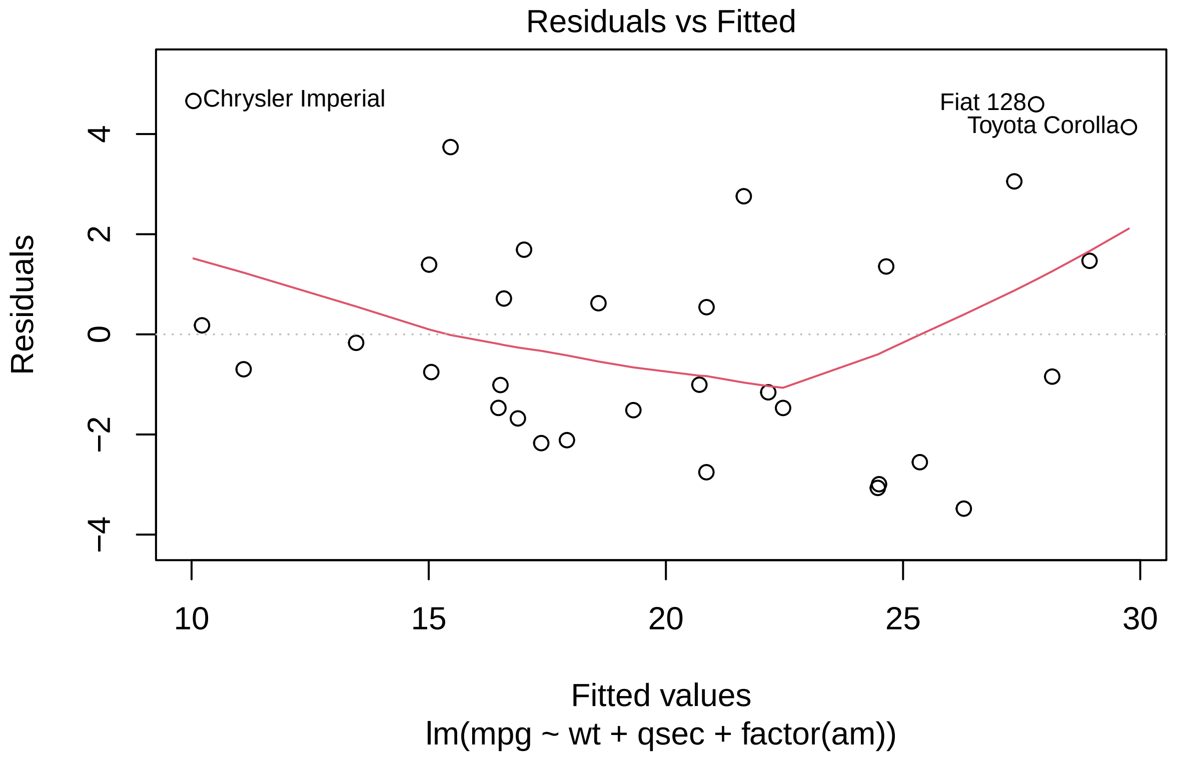
This time looking at variance inflation factors reveal that the numbers are reasonable and we dont detect any sign of collinearity.

```
fitvif <- lm(mpg ~ wt+qsec+factor(am), data = mtcars)
kable(vif(fitvif),align = 'c')
```

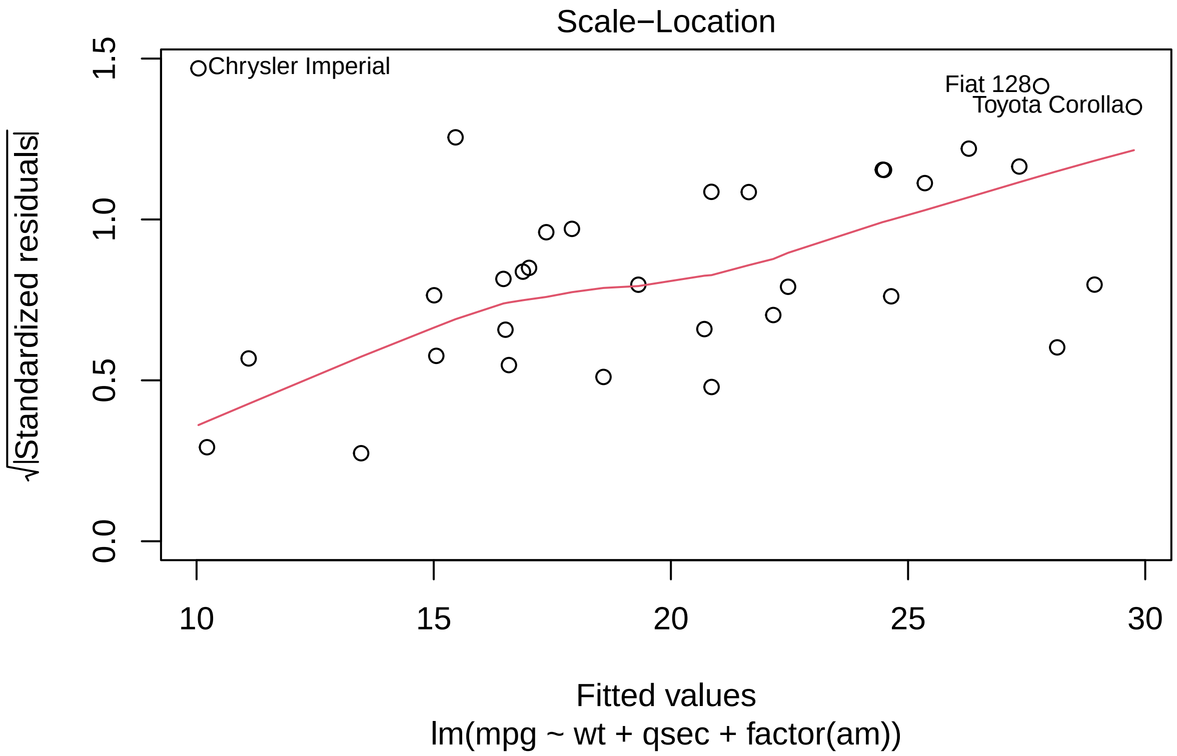
	x
wt	2.482951
qsec	1.364339
factor(am)	2.541437

Residuals versus the fitted values By plotting residuals versus the fitted values, we're looking for any sort of pattern. Same thing with the fitted values versus the standardized, where it's plotting a function of the standardized residuals. Plots below show that no specific pattern exist in the residuals.

```
plot(finalfit, which=1)
```

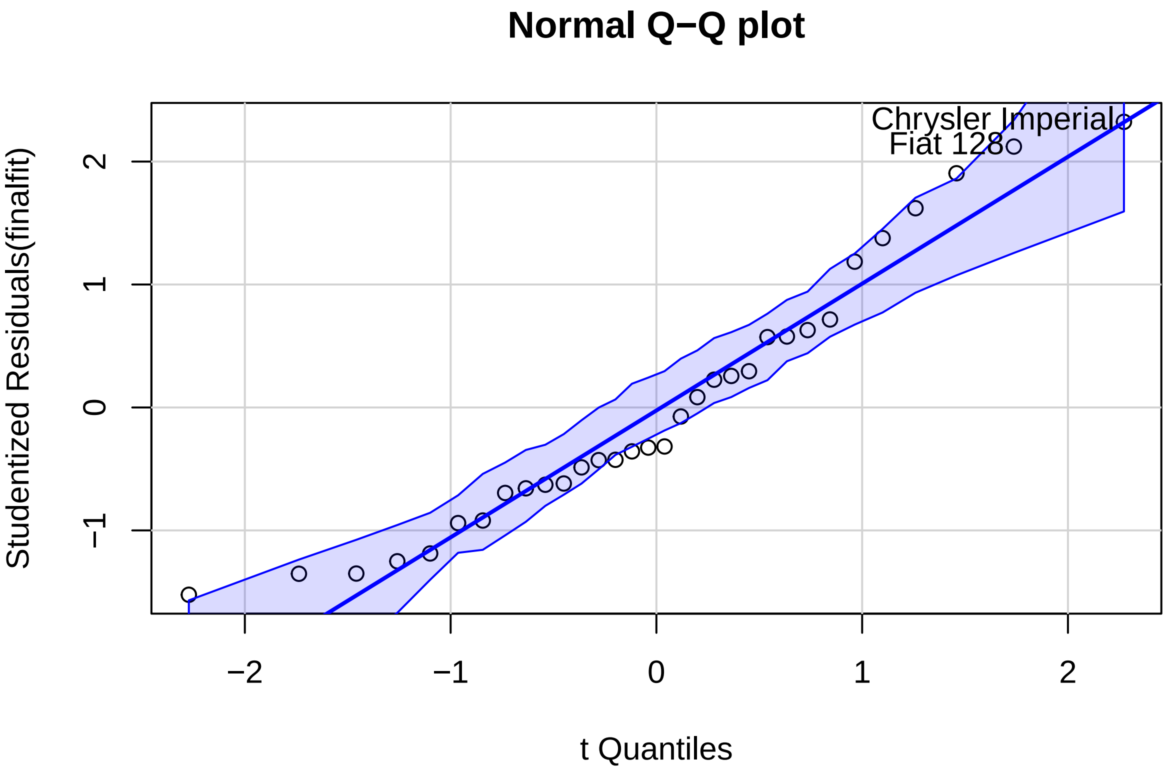


```
plot(finalfit, which=3)
```



Normality of residuals The normal Q-Q plot, is you're trying to figure out the normality of the errors by plotting the theoretical quantiles of the standard normal distribution by the standardized residuals.

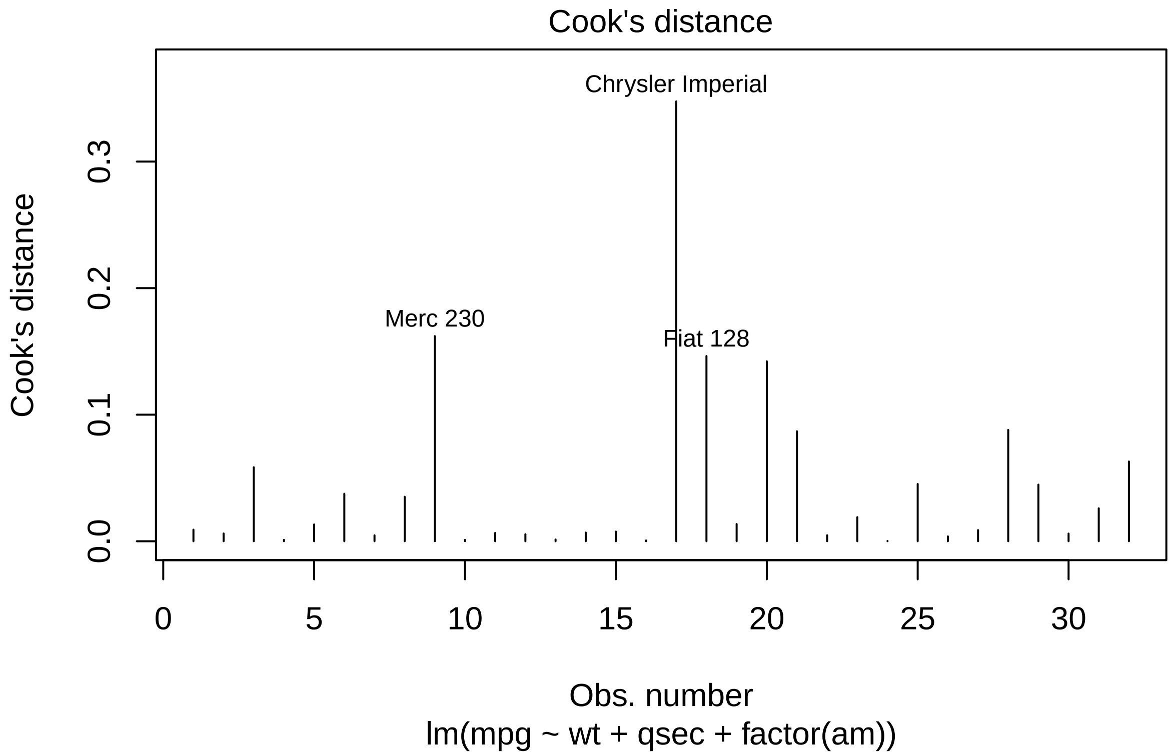
```
qqPlot(finalfit, main="Normal Q-Q plot")
```



```
## Chrysler Imperial          Fiat 128
##                17                18
```

Influential Observations Residuals versus leverage and also cooks distance, that's where we want to look at the comparison of fit at that point verses the potential for influence of that point. So this is also a very useful plot to look at.

```
plot(finalfit, which=4)
```


Cook's distance
Chrysler Imperial
Merc 230
Fiat 128
Cook's distance
0.0
0.1
0.2
0.3
0
5
10
15
20
25
30
Obs. number
lm(mpg ~ wt + qsec + factor(am))