Is an Automatic transmission better than a manual transmission

preethi

11/20/2019

Executive Summary

Here at *MotorTrend*, we want to answer to very simple questions

1) Is an automatic or manual transmission better for MPG
2) Can I quantify the MPG difference between automatic and manual transmissions

We will be using the mtcars dataset to answer these questions

MTCARS dataset

We will begin with describing the mtcars dataset. This data was extracted from the 1974 *MotorTrend* magazine and comprises of fuel consumption and 10 other aspects of automobile design (mpg, number of cylinders, displacement, transmission, engine etc. . .) for 32 automobiles

We can look at the first few rows of this dataset now

```
head(mtcars)
```

```
##                   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
```

From this, we can see that am, the transmission is a binary variable 0 = automatic and 1= manual Now let us get a summary of the variables we are interested in

```
summary(mtcars$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   15.43   19.20   20.09   22.80   33.90
```

```
summary(mtcars$am)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.4062  1.0000  1.0000
```

From the above we can see that the mean miles per gallon is 20.09 and the mean of transmission being 0.40 shows that there are roughly similar number of cars in automatic and manual with slightly more automatic cars (median = 0).

Now let us subset this data

```
automatic <- subset(mtcars, am == 0)
manual <- subset(mtcars, am == 1)

mean(automatic$mpg)
```

```
## [1] 17.14737
```

```
mean(manual$mpg)
```

```
## [1] 24.39231
```

From here we can see that there are 19 observations (cars) are automatic and 13 are manual. We can also see that the mean mpg for the automatic transmission vehicles 17.15mpg and the mean for the manual transmission is 24.39mpg. We have plotted this data in the [figure 1] (#figure_1) in the appendix.

Hypothesis Testing

We can mostly see that MPG is better with a manual transmision compared to an Automatic transmission from that graph. However, there is a bit of overlap and to see if there is truly an effect, we can perform a student t.test to test the null hypothesis (H0) that there is no difference between the MPG for manual and automatic transmission

```
ttest <- t.test(mpg~am,data=mtcars)
ttest
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

As the *p-value* is less than 0.05 (p-value=0.001) we can reject the null hypothesis. Now that we rejected the null hypothesis we can fit some models to see if we can quantify the mpg difference with automatic and manual transmission.

A plot of all the variables against mpg is presented in [figure 2] (#figure_2) of the appendix.

Model Fitting

We can see quite a few variables that look like they may be correlated. so lets first try a linear regression model with mpg as the outcome and just transmission type as a predictor

```
fit1 <- lm(mpg~am, data = mtcars)
summary(fit1)
```

```
##
## 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 ***
## am             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
```

From the above model we can see that the residuals are pretty symmetrical around the mean. The mean MPG of an automatic transmission is 17.14 and with a manual transmission this increases by 7.24 mpg. The p-values are also < 0.05 however the Multiple R-squared value is just 0.34, which suggests that our model explains only 34% of the variability in the data. So let's move on and try to fit all our predictor values with mpg as the outcome variable

```
fit2 <- lm(mpg~., data = mtcars)
summary(fit2)
```

```
##
## 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
```

When we include all the variables as predictor values we see that none of the predictor variables seem to correlated with mpg other than a loose relationship with weight. The Multiple R-squared is now 81% which is pretty high but we have overfitted the model at this stage.

So in R there is a function called step which chooses the possible model in AIC (Akaike information criterion which is an estimator of out of sample prediction error and thereby the relative quality of a model) in a step-wise algorithm

```
fit_3 <- (step(lm(mpg~.,data = mtcars),trace=0))
summary(fit_3)
```

```
##
## Call:
```

```
## lm(formula = mpg ~ wt + qsec + 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 ***
## am            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
```

Conclusion

In accordance with this function, there is a 2.9 mpg increase with manual transmission compared to an automatic transmission and this value is signification ($p < 0.05$). But it is not the variable that shows as an association with mpg. Every 1000lb weight increase will see a decrease of roughly 4 mpg and with every quarter mile time increase shows an increase of about 1.2 mpg

The adjusted R-squared value is 83% which means that a large part of the variation in the data is described by this model.

The residual plots [figure 3] (#figure_3) are given in the appendix and they show that the data is mostly linear (the gentle skewedness could be due to the small sample size).

Appendix

Figure 1

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.1
```

```
mtcars$am <-  as.factor(mtcars$am)
levels(mtcars$am)<- c("Automatic", "Manual")
graph <- ggplot(mtcars,aes(x = am, y = mpg))+geom_boxplot(aes(fill=am))
graph
```

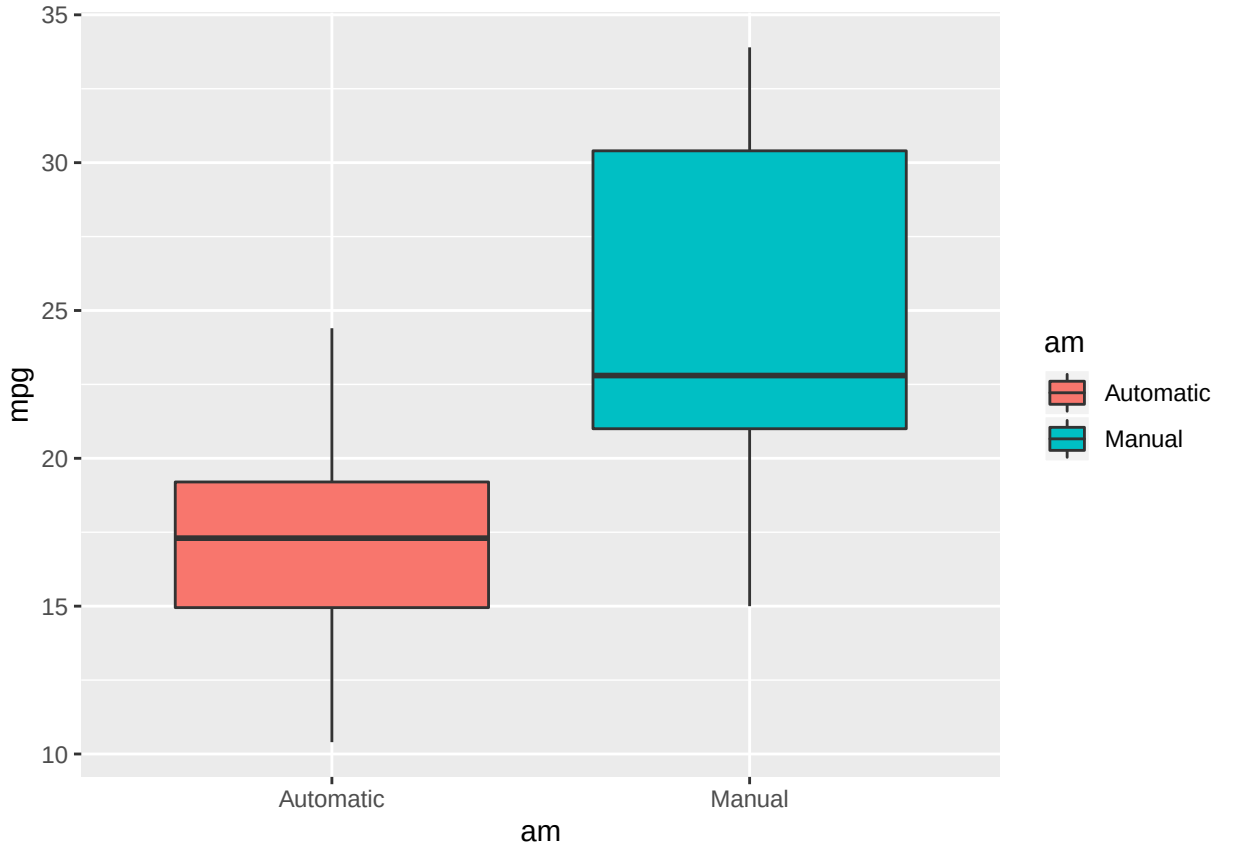


Figure 2

```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 3.6.1
```

```
library(ggplot2)
mtcars %>% gather(-mpg, key = "var", value = "value") %>%
  ggplot (aes(x = value, y = mpg))+ geom_point() +facet_wrap(~ var, scales = "free") + theme_bw()
```

```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```

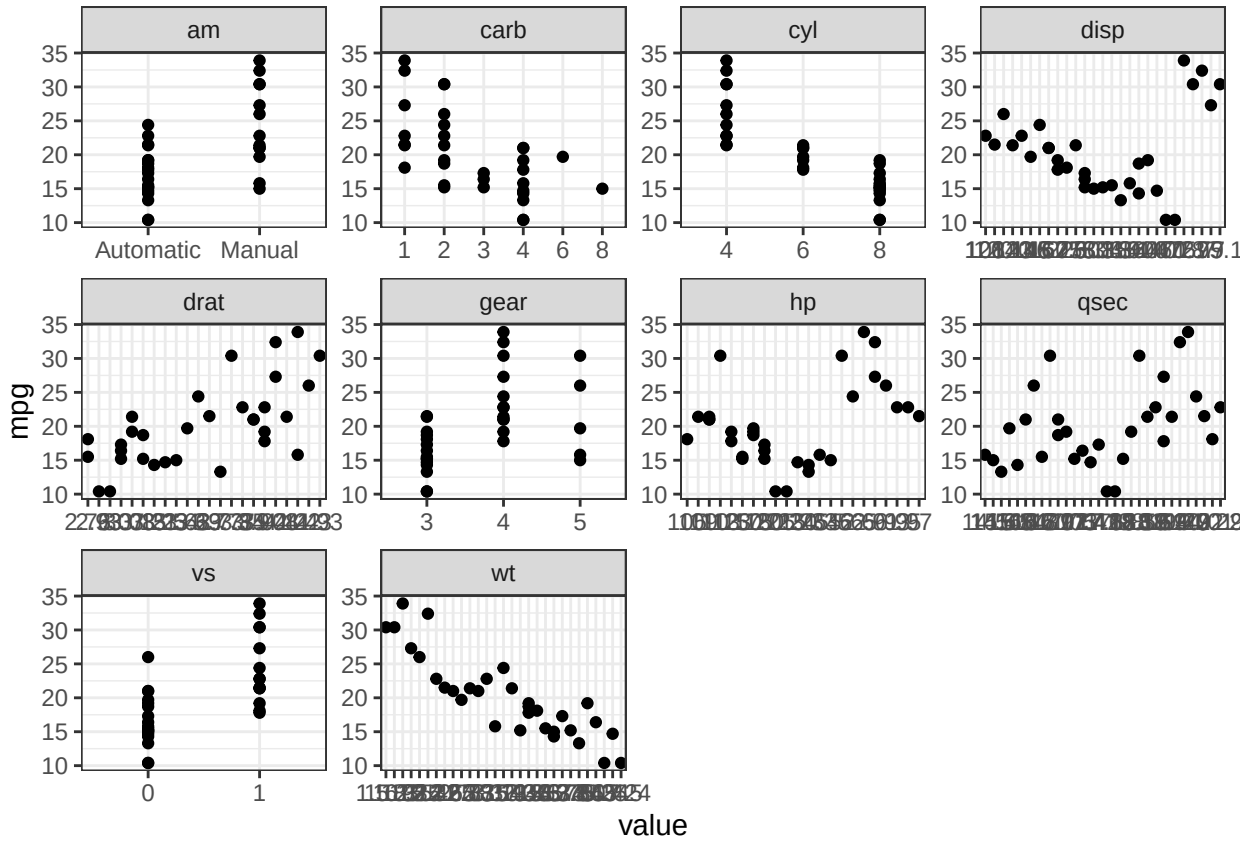
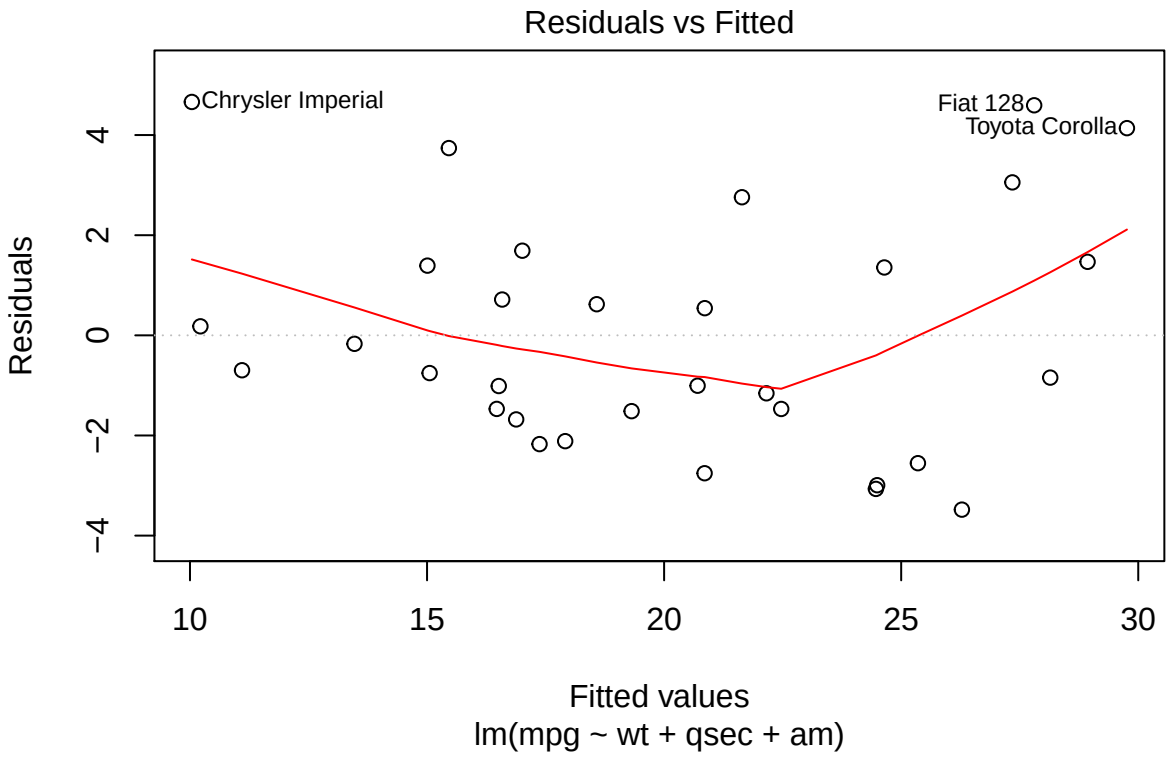
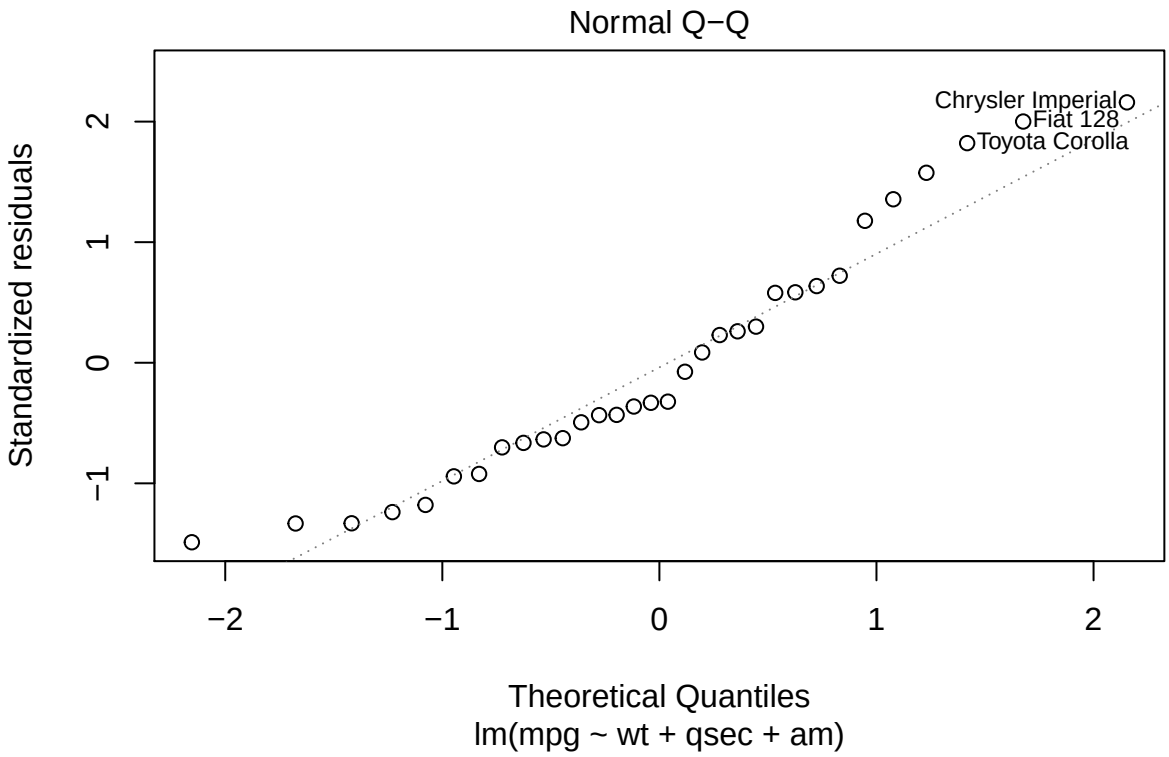
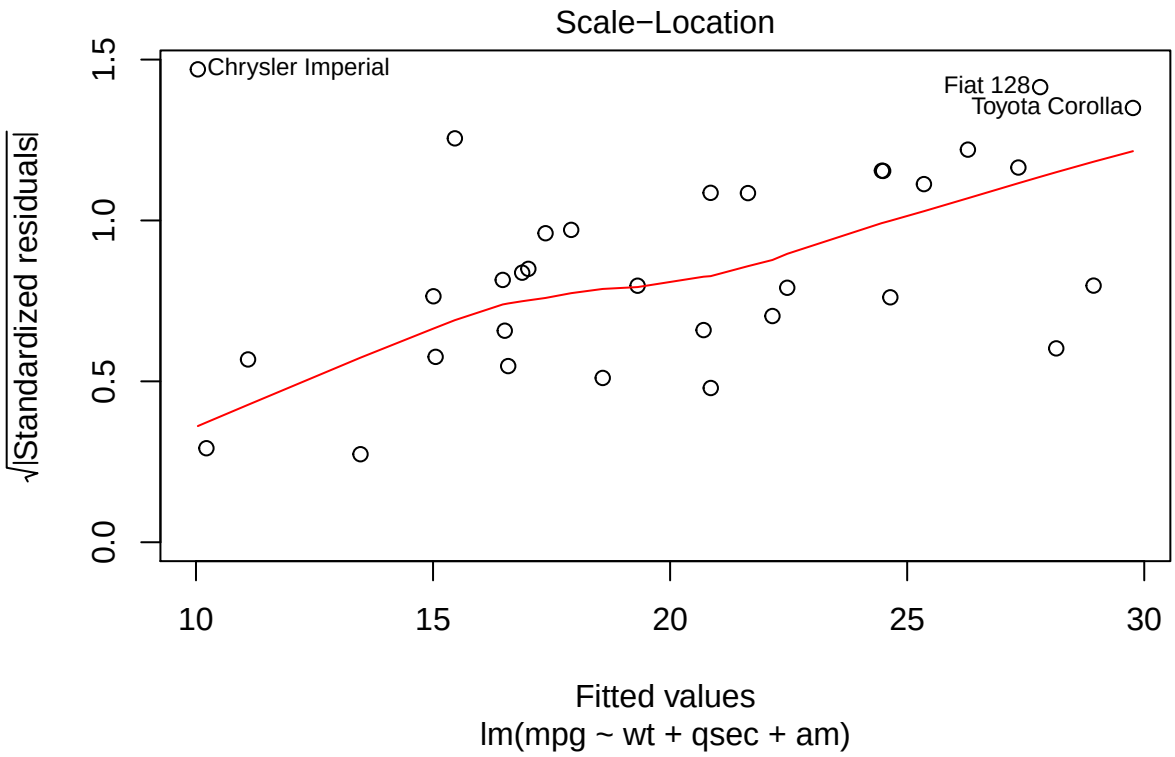


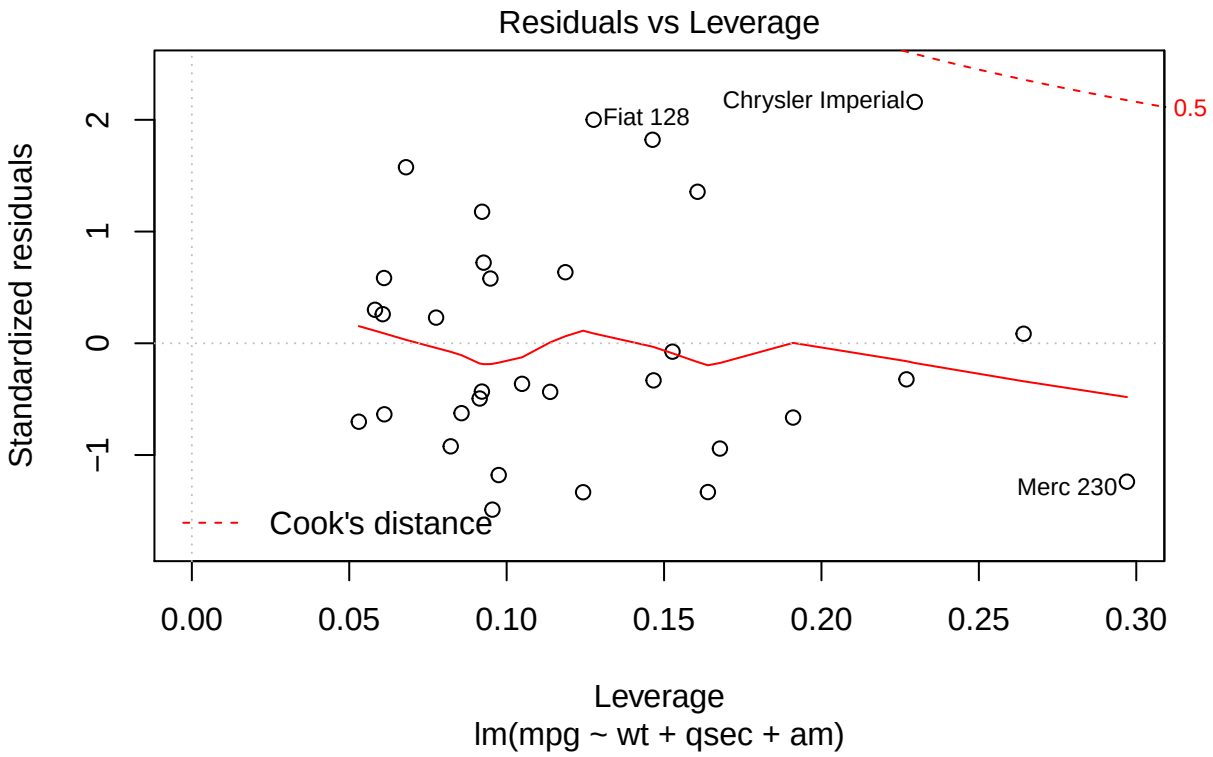
Figure 3

```
plot(fit_3)
```


Residuals vs Fitted
Chrysler Imperial
Fiat 128
Toyota Corolla
Residuals
4
2
0
−2
−4
10
15
20
25
30
Fitted values
lm(mpg ~ wt + qsec + am)


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


Scale−Location
Chrysler Imperial
Fiat 128
Toyota Corolla
√|Standardized residuals|
0.0
0.5
1.0
1.5
10
15
20
25
30
Fitted values
lm(mpg ~ wt + qsec + am)



The first plot, a simple scatter plot between the fitted and the residual values is more or less random except for a few outliers.
The Normal Q-Q plot is a Normal Probability Plot, which gives an almost straight line as the errors are mostly normally distributed. The scale-location plot shows that the residuals are mostly spread equally along the ranges of the predictor. It is homoscedastic. In the Cook's distance we can see that the Ford Pantera and Merc 230 seem to exert some influence.