Regression Models Course Project - Motor Trend

Ajla Dzajic

Executive Summary

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:

1. "Is an automatic or manual transmission better for MPG"?
2. "Quantify the MPG difference between automatic and manual transmissions"

Take the mtcars data set and write up an analysis to answer their question using regression models and exploratory data analyses.

Exploratory Data Analysis

First of all, we'll inspect the data set and see if there are correlations between the regressors. From Figure 1 in appendix we see that many variables in our data set mtcars are highly correlated with MPG, as well as with each other. That means that we'll have to be careful when choosing our regression model because inclusion of highly correlated variables might cause increase in standard errors for other regressors. Second, we'll inspect the impact of the type of transmission on MPG graphically. From Figure 2 in appendix we see that automatic transmission significantly increases consumption of gasoline. To check if that increase is statistically significant we'll do statistical inference t.test. In this case our null hypothesis is that there is no difference in mpg between manual and automatic transmission.

```r
data(mtcars)
mtcars$am <- factor(mtcars$am, labels = c('Automatic', 'Manual'))
t.test(mtcars$mpg[mtcars$am == 'Manual'], mtcars$mpg[mtcars$am == 'Automatic'])
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars$mpg[mtcars$am == "Manual"] and mtcars$mpg[mtcars$am == "Automatic"]
## 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:
##   3.209684 11.280194
## sample estimates:
## mean of x mean of y
##  24.39231  17.14737
```

As we can see from the output our p-value is 0.001374 which is less than 0.05. We will reject the null hypothesis and conclude that there is a difference in mpg between automatic and manual transmission. We conclude that there is a difference in mpg between automatic and manual transmission and that the difference is between 3.209684 and 11.280194 mpg.

Of course this test compeletely ignores other factors that might be influencing gasoline consumption. The fact that we didn't take into account correaltions between variables makes us a bit suspicious of our t.test results. For that reason, we'll proceed with further analysis.

Regression Analysis

First I used backward elimination where I started with the model that includes all variables in mtcars as predictors and gradually removed those whose p-values were highest until all the remaining variables were significant predictors. Using this method I came to conclusion that the best fitted model would be one that includes wt - Weight (lb/1000), qsec - 1/4 mile time and am -Transmission as predictors.

```
summary(fit7)
```

```
##
## 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 ***
## amManual      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
```

Using nested likelihood ratio test for model selection

```
fit_am <- lm(mpg ~ am, data = mtcars)
fit <- lm(mpg ~ ., data = mtcars)
anova(fit_am, fit7, fit)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + qsec + am
## Model 3: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 169.29  2    551.61 39.2687 8.025e-08 ***
## 3     21 147.49  7     21.79  0.4432    0.8636
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From our analysis of variance we can see that adding wt and qsec as regressors to our model that consisted only of am as a predictor of mpg improved our model. We would conclude that all of the added Model 2 terms are necessary over Model 1. On the other hand, we see that Model 3 terms are not necesary over Model 2 terms and therefore we would choose Model 2.

We can see that p-value for amManual, an estimate of difference between the automatic and manual transmission, is 0.046716 which is less than 0.05. This means that the difference between the two is statistically significant and we can reject the null hypothesis once more that there is no difference in impact on mpg between automatic and manual transmission.

```
t <- qt(1 - 0.05/2, length(mtcars$am) - 2)
2.9358 + c(-1, 1) * t * 1.4109
```

```
## [1] 0.05435779 5.81724221
```

We can see that the difference in mpg between automatic and manual transmission is between 0.05435779 and 5.81724221.

Appendix

Figure 1

```r
library(GGally) # collection of additional tools for ggplot2
library(ggplot2)
g <- ggpairs(mtcars, lower = list(continuous = 'smooth'))
g
```

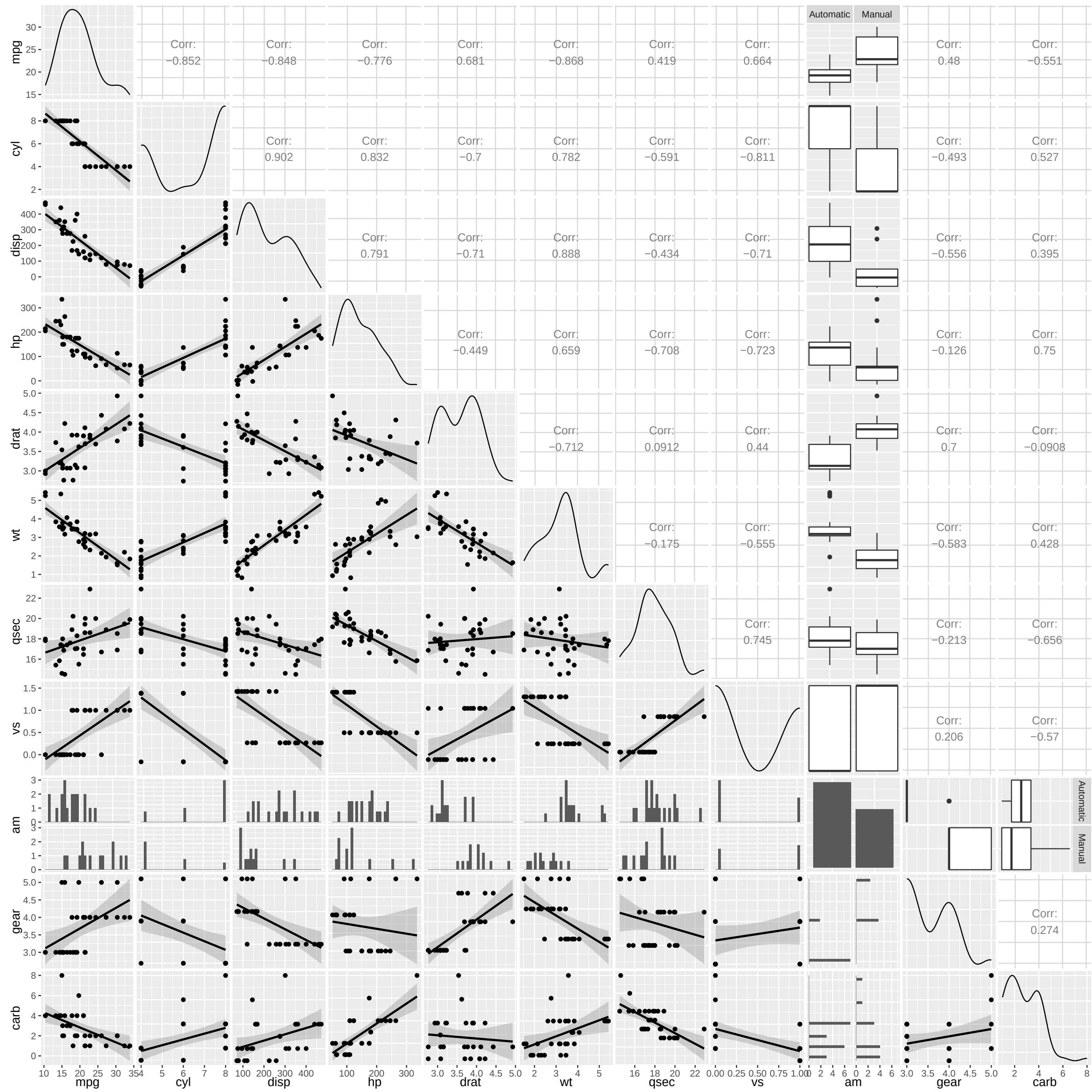


Figure 2

```r
g <- ggplot(mtcars, aes(factor(am), mpg))
g <- g + geom_boxplot(aes(fill = factor(am)))
g <- g + labs(title = 'Effect of Transmission Type on Miles per Gallon',
              x = 'Transmission', y = 'Miles per Gallon')
g
```

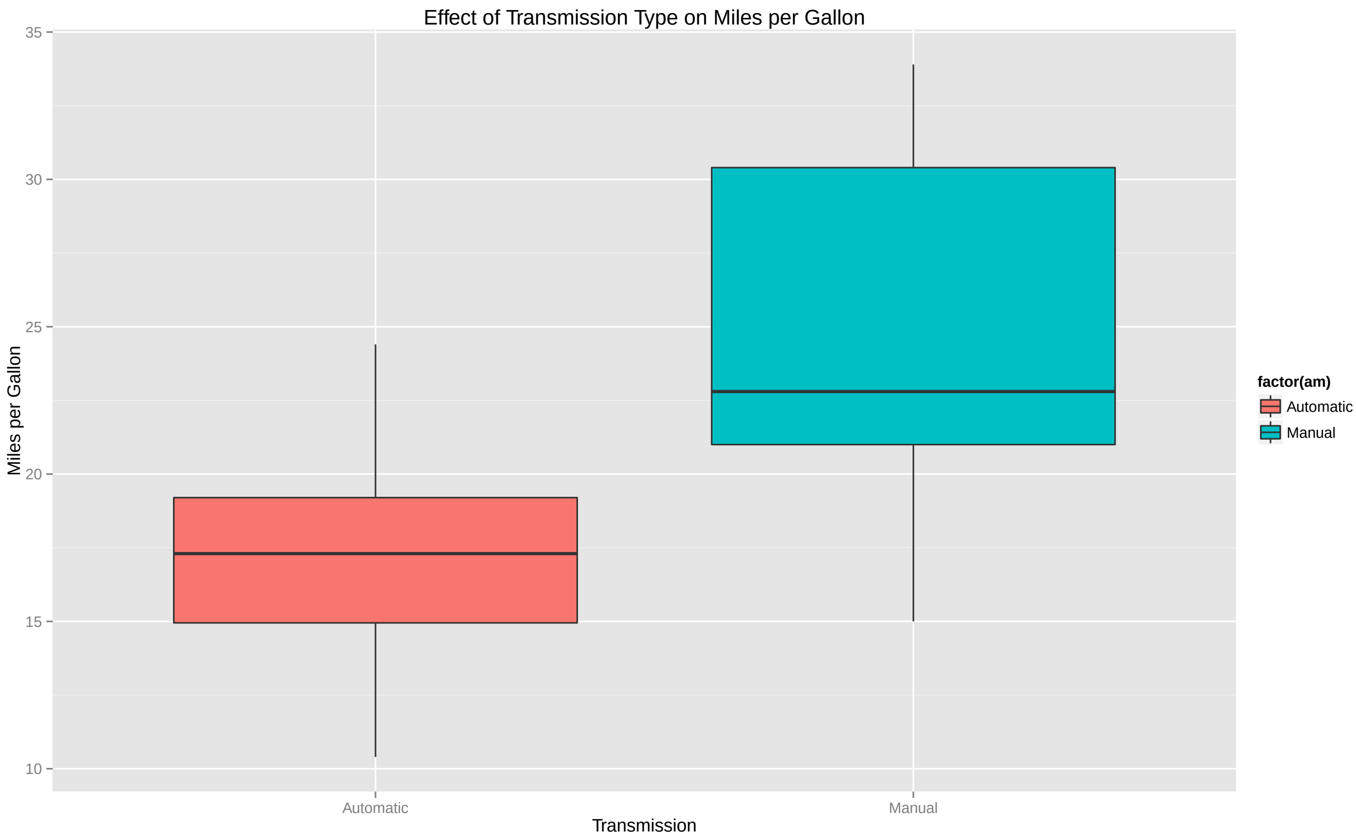
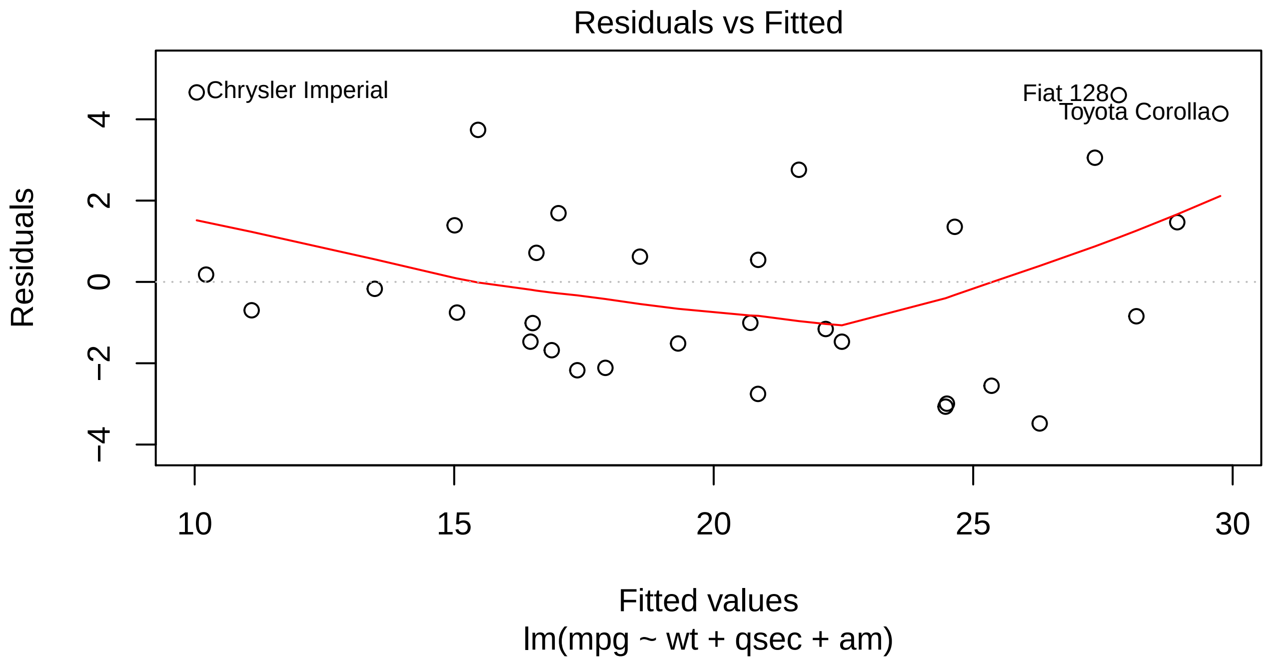
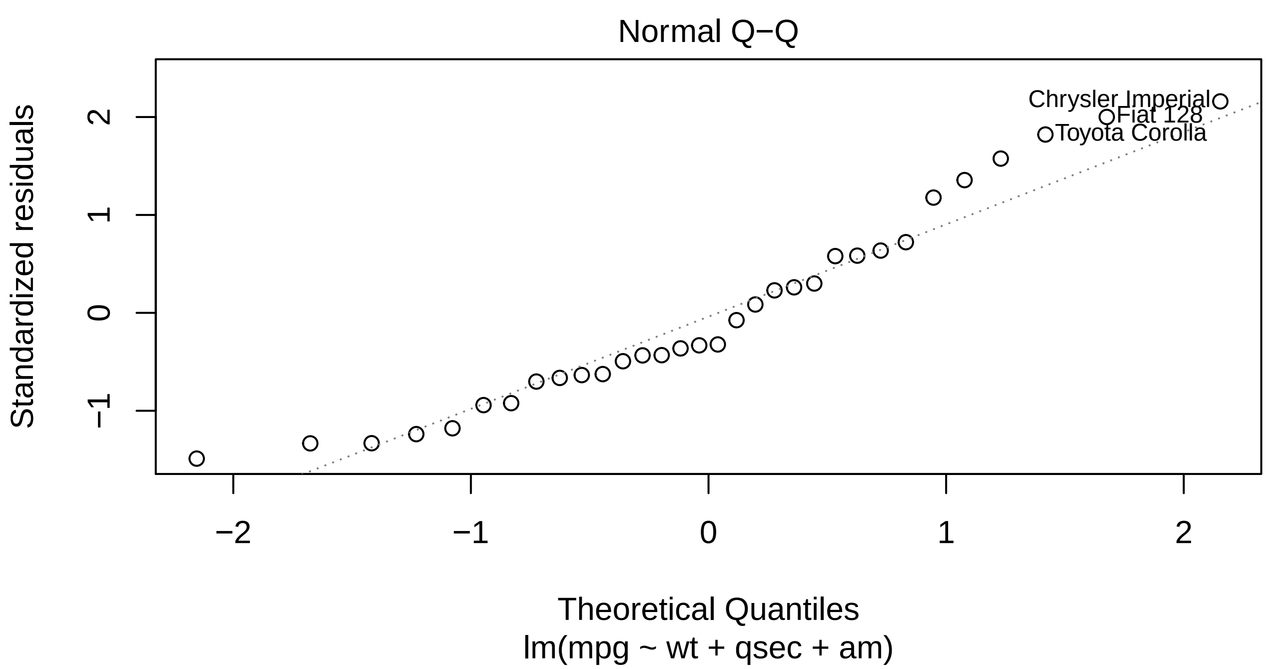
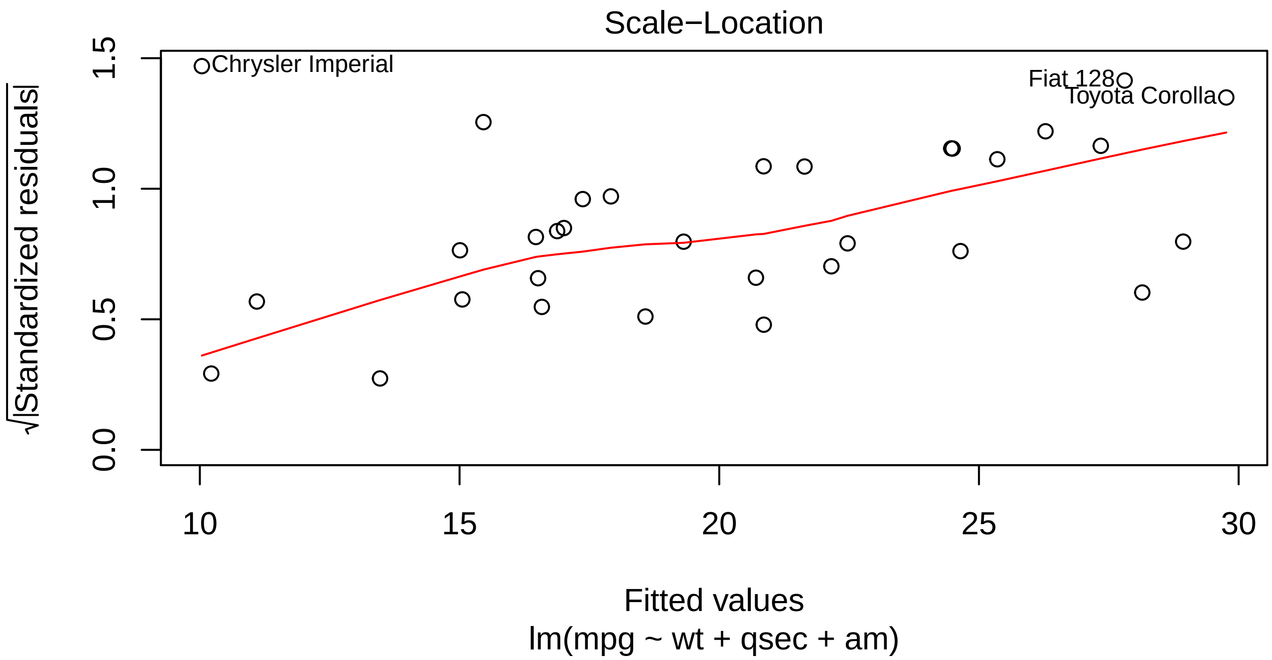


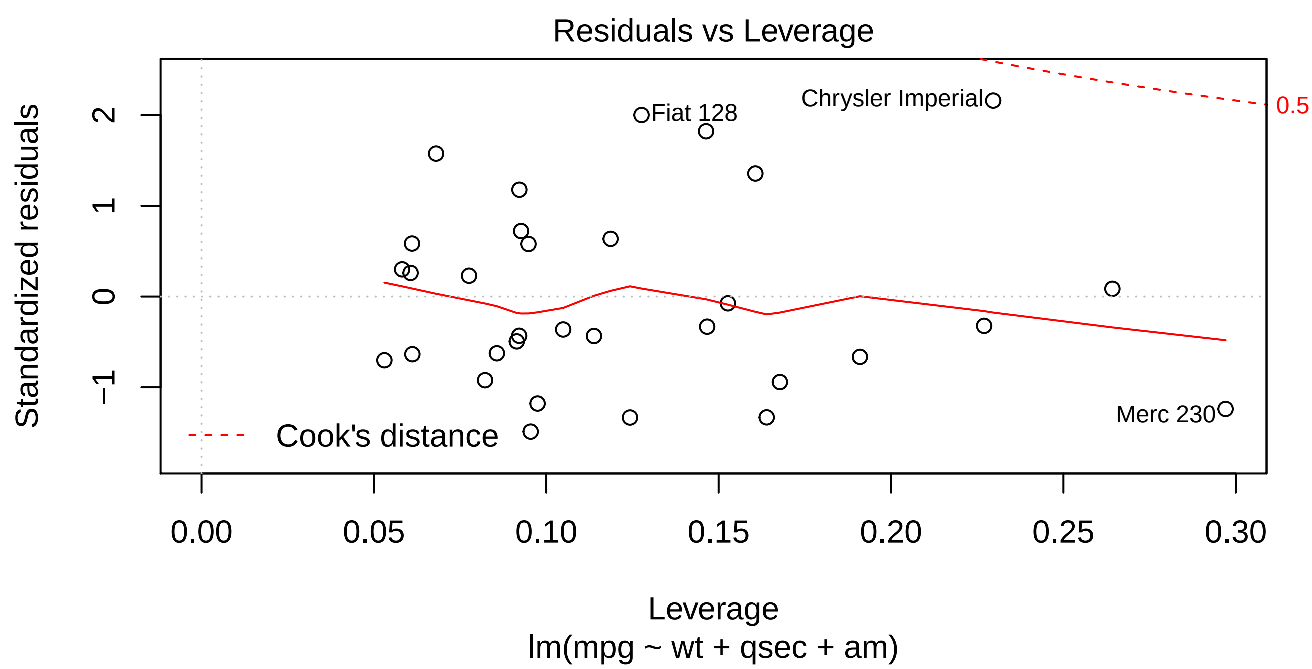
Figure 3 - Residuals plot

```r
plot(fit7)
```






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)


Residuals vs Leverage
Fiat 128
Chrysler Imperial
0.5
Merc 230
Cook's distance
Standardized residuals
−1
0
1
2
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Leverage
lm(mpg ~ wt + qsec + am)