Is an automatic or manual transmission better for MPG?

Lulu Cao

8/2/2020

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:

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

The results of this analysis confirms that the manual transmission should be better than automatic tranmission for MPG. However, with adjustment of other variables, in this case, wt and qsec, the MPG difference between automatic and manual transmissions are greatly reduced from 7.245 to 2.9358.

Data Structure and Processing

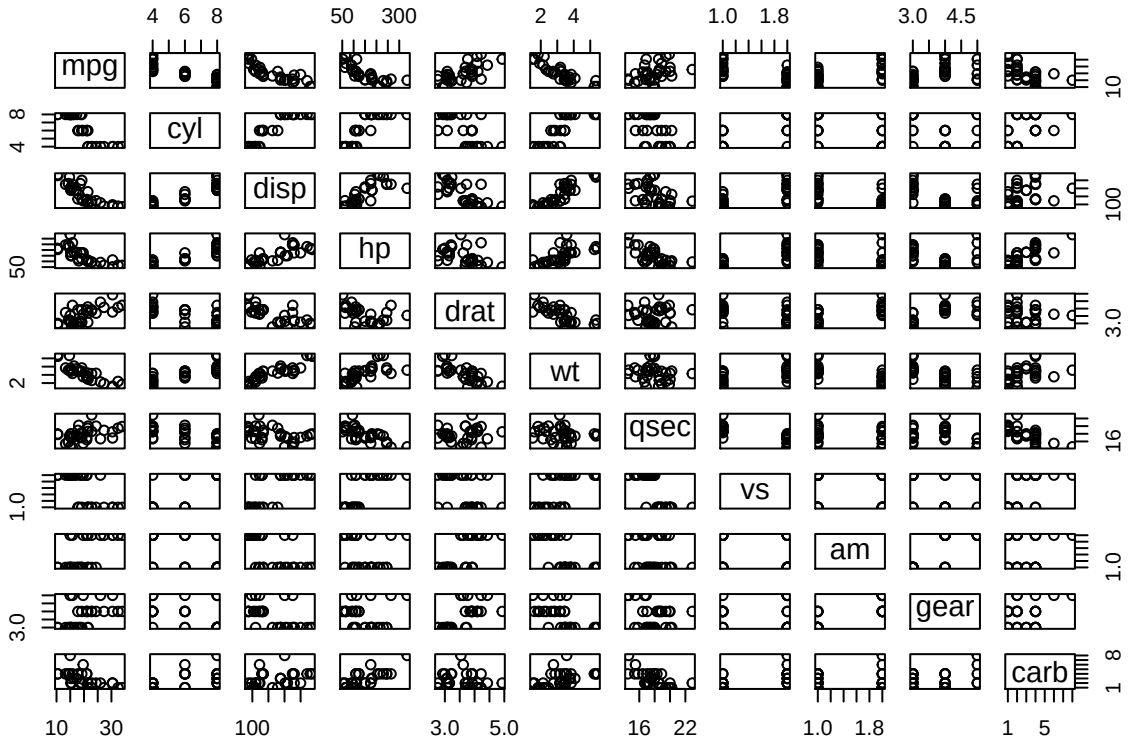
There are total 32 observations with 11 variables.

```
data(mtcars)
mtcars$am <- ifelse(test=mtcars$am==0, yes="auto", no="manual")
mtcars$am <- as.factor(mtcars$am)
mtcars$vs <- ifelse(test=mtcars$vs==0, yes="vshaped", no="straight")
mtcars$vs <- as.factor(mtcars$vs)
```

Exploratory Analysis

Scatter plot of 11 variables

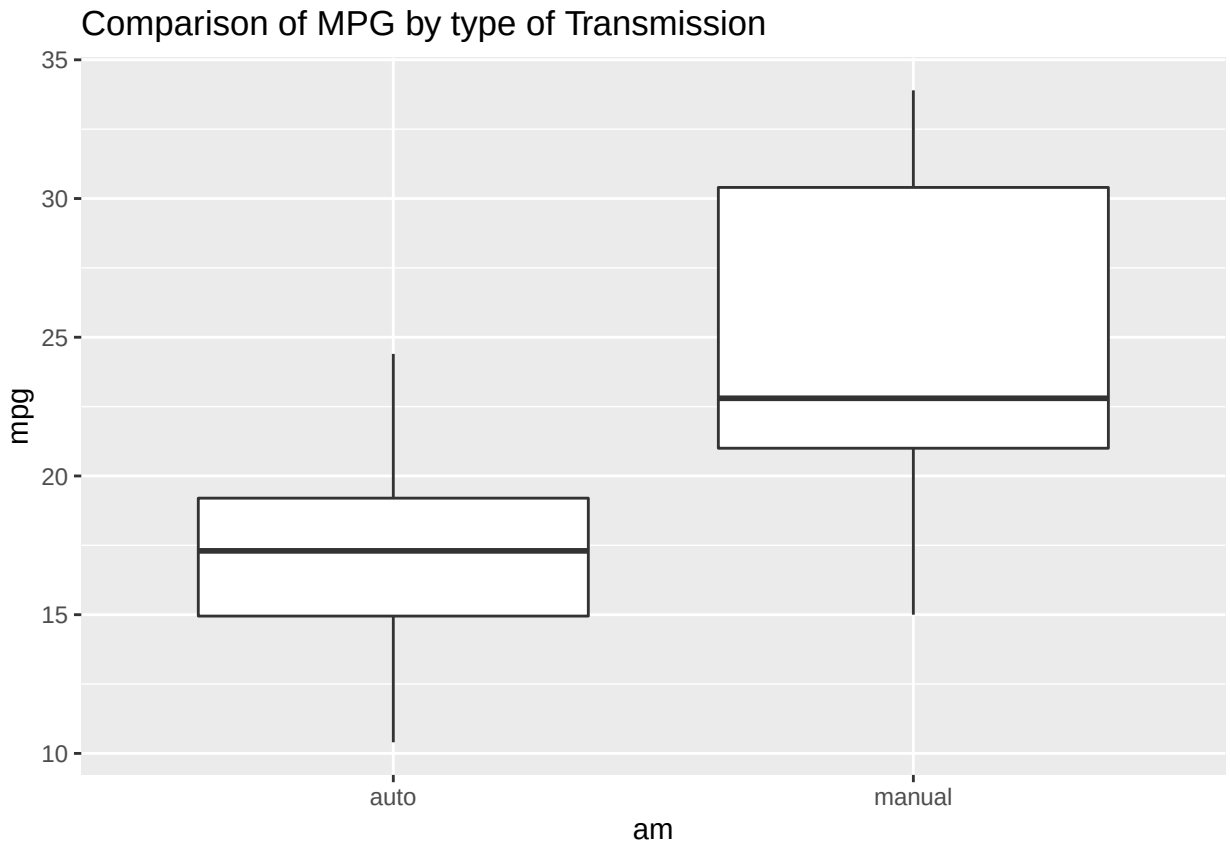
```
plot(mtcars)
```



All 11 variables are generally even distributed.MPG appears to be correlated with all 11 variables. Many of them are highly correlated with each other. A multivariable linear regression model needs to be carefully built to analyze the impact of transmission on MPG.

Make Boxplot to compare mpg between auto and manual transmission

```
library(ggplot2)
g1 <- ggplot(mtcars, aes(x=am, y=mpg))+geom_boxplot()+labs(title="Comparison of MPG by type of Transmiss
g1
```



The boxplot shows that cars with manual tranmission appears to have better mpg compared to cars with auto transmission. However, we need to further look at the confunders and verify the correlations.

Single linear Regression Models

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

Using am as the only regressor, cars with manual transmission is signficantly more efficient compared to cars with auto transmission. On average, 7.245 more mpg is associated with manual transmission compared to auto transmission. However, the R-squared for this model is only 0.3385, and it suggests that the model is poorly built. We need to look at additional variables for a better fit and find out the confunding factors for comparing mpg between manual and auto transmission.

Build Multiple Linear Regression Models

First, fit mpg against all 11 variables

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

Adjusted R-squared is 0.8066, and residual standard error is 2.65. P values for cyl, drat, vs, gear, and carb are relative larger compared to others. Removing these 5 variables and try the next fit.

```
fit2 <- lm(mpg~disp+hp+wt+qsec+am, data=mtcars)
```

Adjusted R-squared is 0.8375, and residual standard error is 2.429. P values for disp and hp are larger than 0.05. Test removing disp and hp from the model.

```
fit3 <- lm(mpg~wt+qsec+am, data=mtcars)
```

Adjusted R-squared is 0.8336, and residual standard error is 2.459. Removing disp and hp simplies the model and R-squared is not impacted much.

```
fit4 <- lm(mpg~wt+am, data=mtcars)
fit5 <- lm(mpg~qsec+am,data=mtcars)
anova(fit,fit4,fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + am
## Model 3: mpg ~ wt + qsec + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 73.203 2.673e-09 ***
## 3     28 169.29  1    109.03 18.034 0.0002162 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit,fit5,fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ qsec + am
## Model 3: mpg ~ wt + qsec + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 352.63  1    368.26 60.911 1.679e-08 ***
## 3     28 169.29  1    183.35 30.326 6.953e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion

Anova analysis suggests that including wt and qsec in the model significantly improves the fit. Combining all the analysis results, fit 3 is the simpliest model that gives the largest adjusted R-squared (0.8336) and the least residual standard error (2.459). In fit 3, the coef for ammanual is 2.9358 and this means the average mpg for manual transmission is 2.9358 higher compared to the auto transmission. The p value is 0.046716 and therefore the difference is statistically significant ($p < 0.05$).

Appendix

```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  : Factor w/ 2 levels "straight","vshaped": 2 2 1 1 2 1 2 1 1 1 ...
##  $ am  : Factor w/ 2 levels "auto","manual": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```r
summary(fit)
```

```
##
## 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 ***
## ammanual       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
```

```r
summary(fit1)
```

```
##
## 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) 12.62114   19.02842   0.663   0.5144
```

```
## 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
## vsvshaped   -0.31776    2.10451  -0.151   0.8814
## ammanual     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
```

```
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ disp + hp + wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5399 -1.7398 -0.3196  1.1676  4.5534
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.36190    9.74079   1.474  0.15238
## disp         0.01124    0.01060   1.060  0.29897
## hp          -0.02117    0.01450  -1.460  0.15639
## wt          -4.08433    1.19410  -3.420  0.00208 **
## qsec         1.00690    0.47543   2.118  0.04391 *
## ammanual     3.47045    1.48578   2.336  0.02749 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.429 on 26 degrees of freedom
## Multiple R-squared:  0.8637, Adjusted R-squared:  0.8375
## F-statistic: 32.96 on 5 and 26 DF,  p-value: 1.844e-10
```

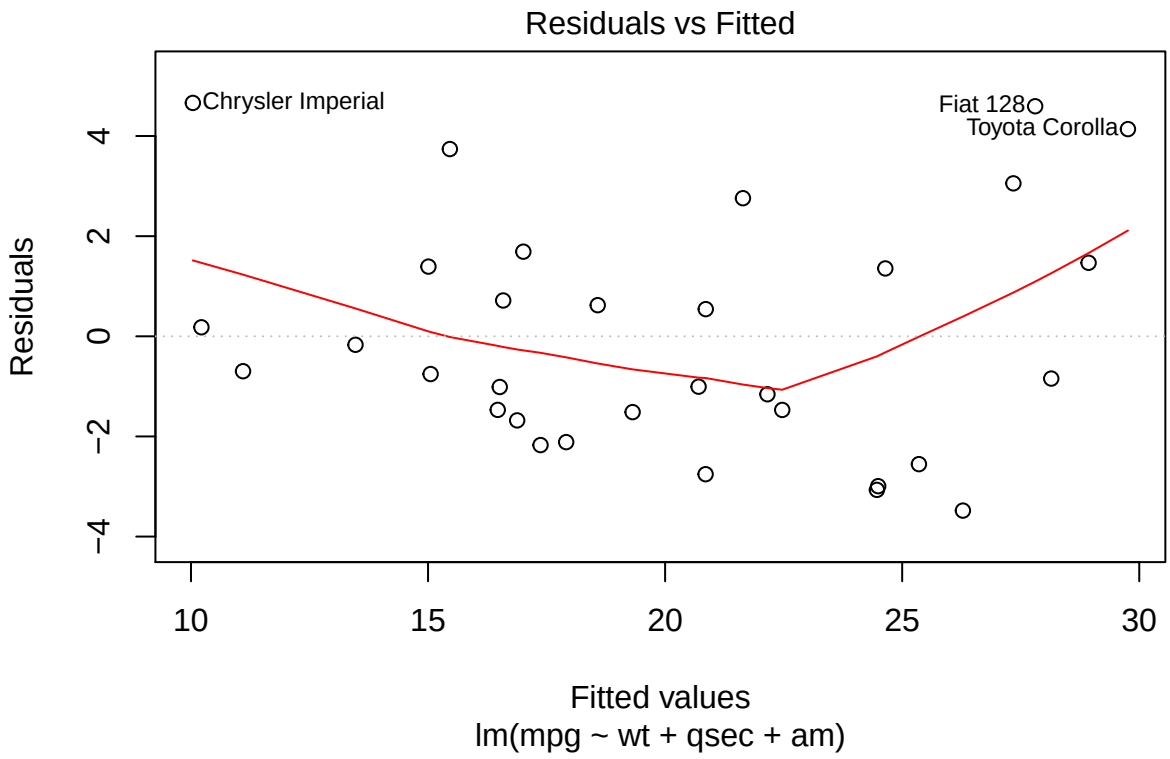
```
summary(fit3)
```

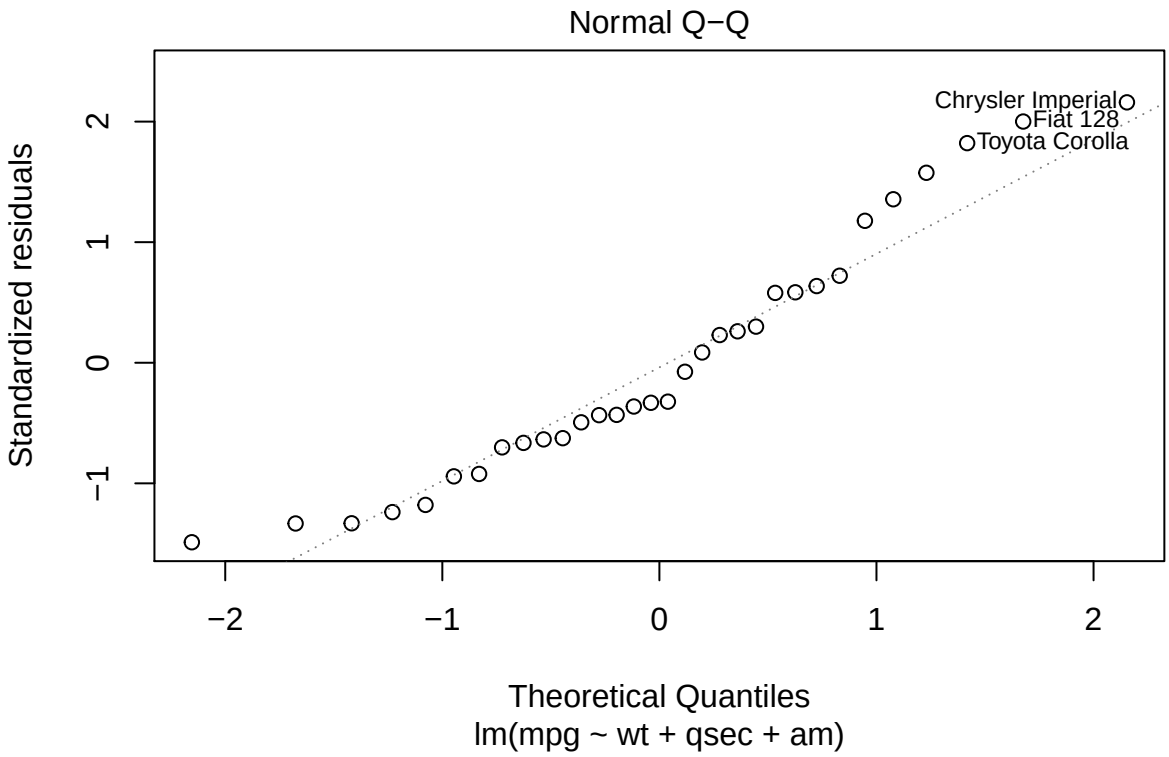
```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
```

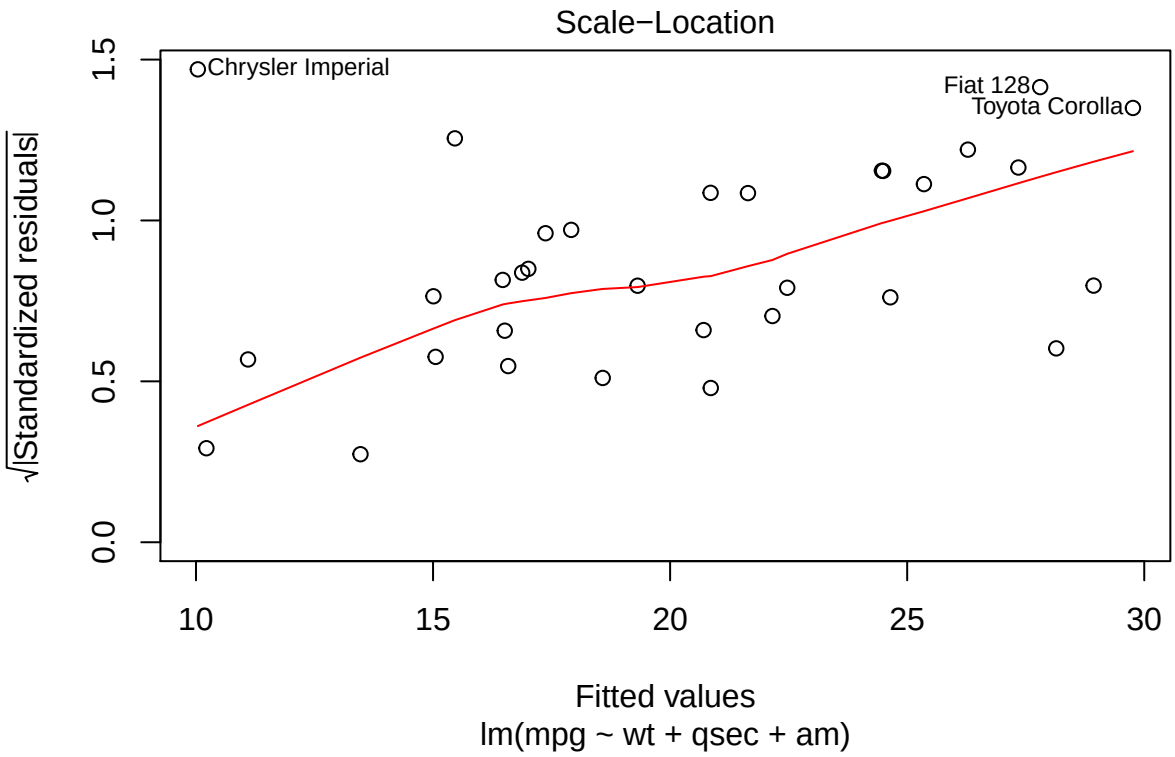
```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## ammanual      2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 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
```

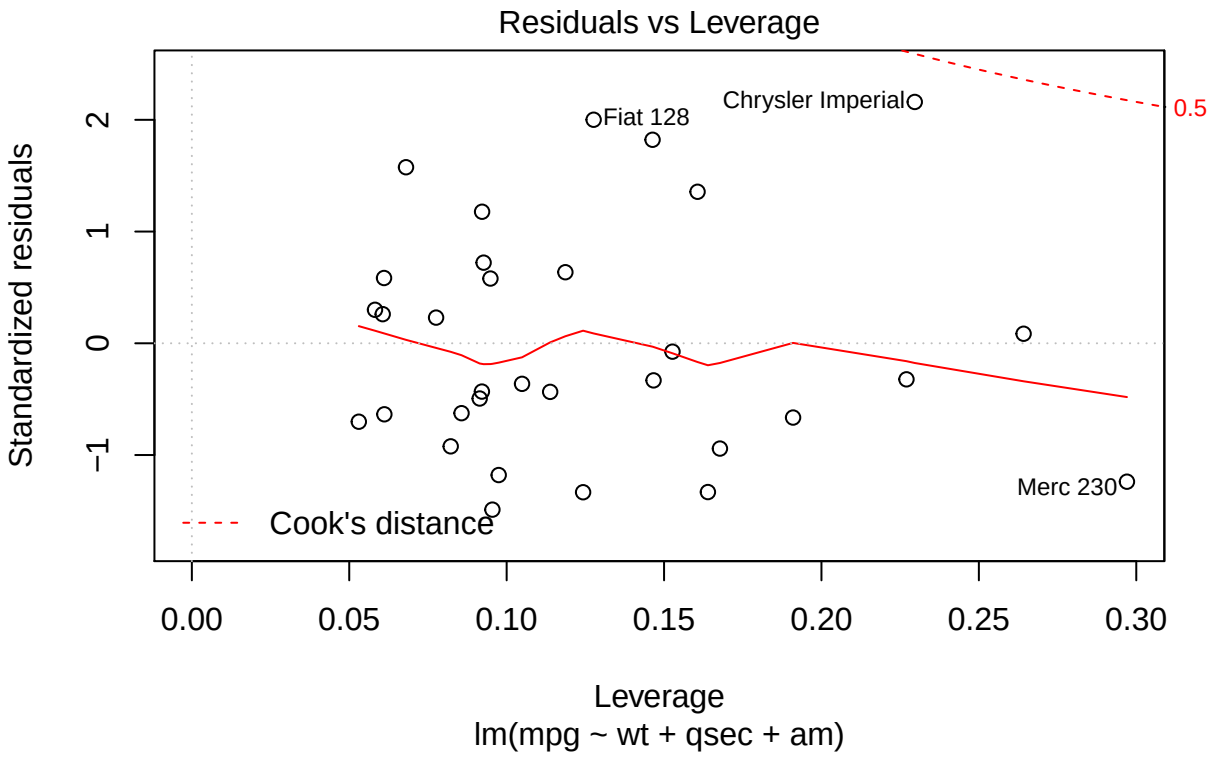
Diagnosis of fit3

```
plot(fit3)
```




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)



Visulization of the residual plots of fit3 shows no systematic pattern, suggesting the model fit3 is reasonablely good.