Motor Trend analysis

Izaak Jephson

17/03/2020

Executive Summary

An analysis was carried out on the mtcars data set to determine whether automatic or manual transmission was better for mpg. The analysis showed that it appears that maunal transmission is associated with an increased mpg over automatic transmission. However, when the confounding variables of horsepower and weigth are included within the model, the variation in transmission does not appear to contribute to the model significantly.

Data used

The data used here comes from the "mtcars" data included as one of the data sets in the "datasets" package in R. The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). The variables of interest in this analysis are "mpg" containing the miles per (US) gallon and "am" which contains the transmission mode for the vehicle (0 = automatic, 1 = manual). Also considered were the variables "wt" (weight in 1000 lbs), "hp" (gross horsepower) and "cyl" (number of cylinders).

Exploratory data analysis

```
data(mtcars)
```

First we simply plot the data (shown in appendix). From the violin plot, there does seem to be a difference in mpg for automatic and manual transmission. Automatic transmission vehicles have a mean mpg of 17.1, while manual vehicles have a mean mpg of 24.4. The variance of mpg for the manual transmission vehicles is larger at 38, compared to 14.7 for the automatic transmission vehicles. This is summarized below:

```
mtcars %>%
        group_by(am) %>%
        summarise(
                mean = mean(mpg) %>% round(digits = 1),
                variance = var(mpg) %>% round(digits = 1)
        ) %>%
        mutate(am = case_when(am == 0 ~ "automatic",
                              am == 1 ~ "manual")) %>%
        rename(transmission = am) %>%
        kable(caption = "Mean and variance of mpg by mode of transmission")
```

Table 1: Mean and variance of mpg by mode of transmission

transmission	mean	variance
automatic	17.1	14.7
manual	24.4	38.0

We can test to see if this difference is significant:

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

```
##
##  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
```

The 95% confidence interval for the difference in the means of mpg for the two transmission modes is -11.3 to -3.2. This does not include zero, therefore, we can reject the null hypothesis that there is no difference between the two modes of transmission at the 95% confidence level. Alternatively we could use the p value of 0.0014 and conclude that we have eveidence for a difference in mpg for the two groups at the 95% confidence level as the p value is below 0.05.

Regression

Linear Model

Now that we have established that there is a difference between the two modes of transmission, we can attempt to quantify this difference using regeression. We can start by fitting a simple linear model to the data, using only am as a predictor for mpg.

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

Fitting a linear model gives a slope of 7.2449 and intercept of 17.147. This can be intepreted as indicating that manual cars have mpg of 7.2 higher than automatics. This agrees with our earlier exploratory data analysis as expected (it is the difference of the means).

From the graph, we can see that the linear model is a reasonable fit for the data. We can see this further by looking at a plot of the residuals (see appendix), which we would expect to be normally distributed with mean zero, if the model were a good fit.

As can be seen in the residuals plot, the points are relatively evenly distributed above and below zero and we can conclude that this model is a reasonable fit to the data.

Multivariate analysis

It is likely that other variables are correlated with both mpg and transmission. We should check that the variation in mpg with transmission is not explained by other factors. To do this, we can analyse multiple nested models using ANOVA. We start by looking at horsepoer, weight, cylinders and transmission as potential influencing variables, then look at a model which exlcudes transmission as a variable.

```
model_1 <- lm(mpg ~ hp + wt + cyl + am, data = mtcars)
model_2 <- lm(mpg ~ hp + wt + cyl, data = mtcars)
model_3 <- lm(mpg ~ hp + wt, data = mtcars)
model_4 <- lm(mpg ~ hp, data = mtcars)

anova(model_1, model_2, model_3, model_4)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ hp + wt + cyl + am
## Model 2: mpg ~ hp + wt + cyl
## Model 3: mpg ~ hp + wt
## Model 4: mpg ~ hp
##   Res.Df    RSS Df Sum of Sq       F   Pr(>F)
## 1     27 170.00
## 2     28 176.62 -1    -6.623  1.0519  0.31418
## 3     29 195.05 -1   -18.427  2.9267  0.09859 .
## 4     30 447.67 -1  -252.627 40.1236 8.81e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

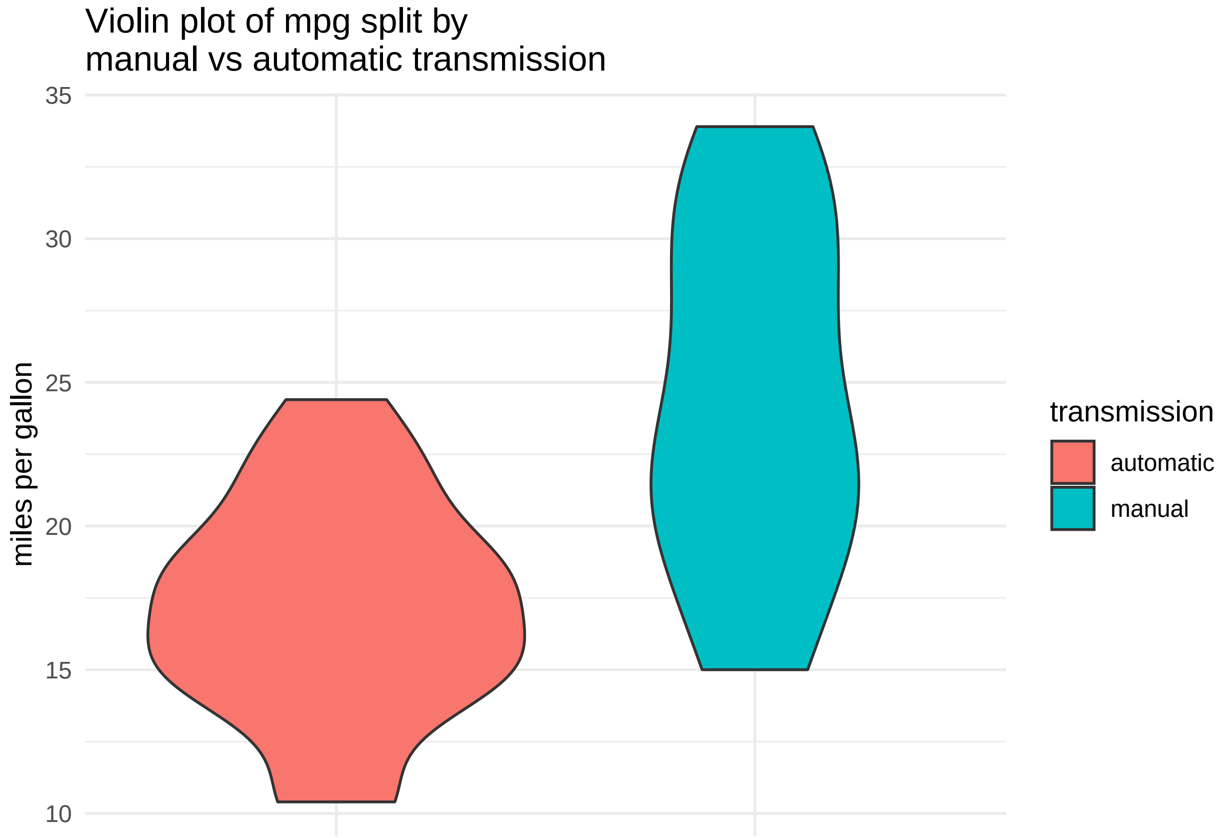
This analysis appears to suggest that the models including transmission and cylinders are not significantly different to the model including only horsepower and weight. This implies that the difference in mpg with varying transmission is explained by the difference in weight and horsepower (though the causation could also run the other way, ie the variation in transmissions causes differences in horsepower, weight and mpg).

We can also check the residuals of the fits, shown in the appendix. These all look reasonable and the residuals appear to be approximately normally distributed. Although there are some clear outliers in the horsepower only model, which implies some of the variation is being missed here, as suggested by the ANOVA.

Appendix

Violin plot of mpg by transmission

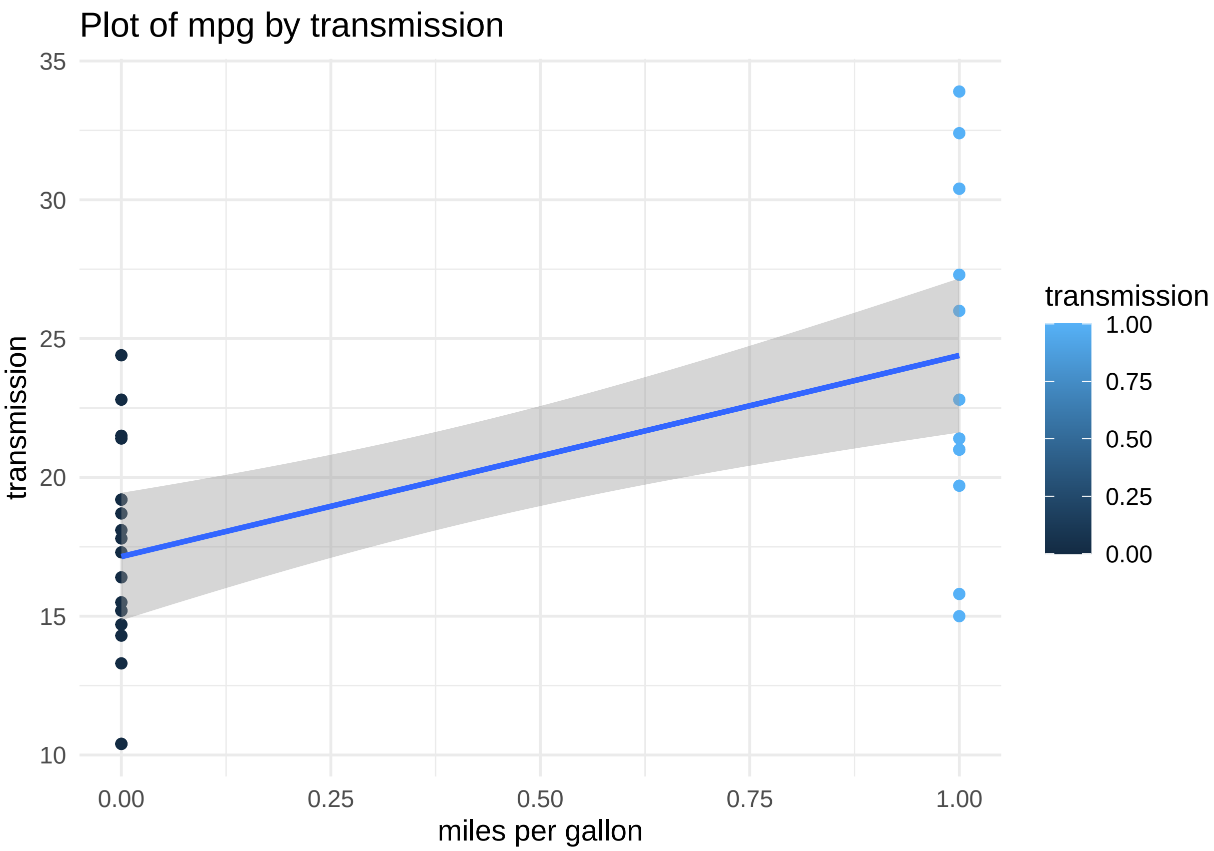
```
mtcars %>%
        ggplot(aes(x = factor(am),
                   y = mpg,
                   fill = factor(am))) +
        geom_violin() +
        theme_minimal() +
        labs(title = "Violin plot of mpg split by\nmanual vs automatic transmission",
             x = "transmission",
             y = "miles per gallon") +
        scale_fill_discrete(name = "transmission",
                            labels = c("automatic", "manual")) +
        theme(
                axis.title.x = element_blank(),
                axis.text.x = element_blank(),
                axis.ticks.x = element_blank()
        )
```



Plot of mpg by transmission with linear model fitted

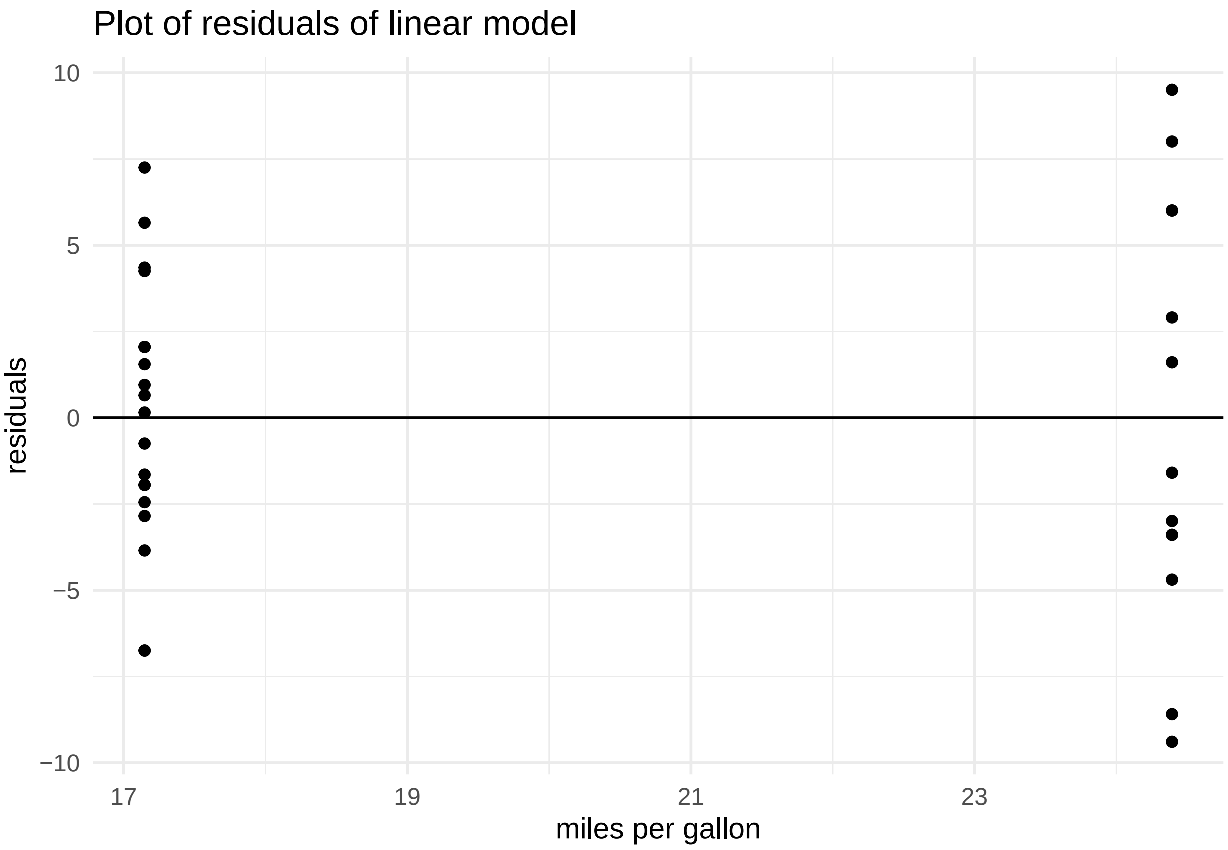
```
mtcars %>%
        ggplot(aes(y = mpg,
                   x = am,
                   colour = am)) +
        geom_point() +
        theme_minimal() +
        labs(title = "Plot of mpg by transmission",
             y = "transmission",
             x = "miles per gallon") +
        scale_colour_continuous(name = "transmission") +
        geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



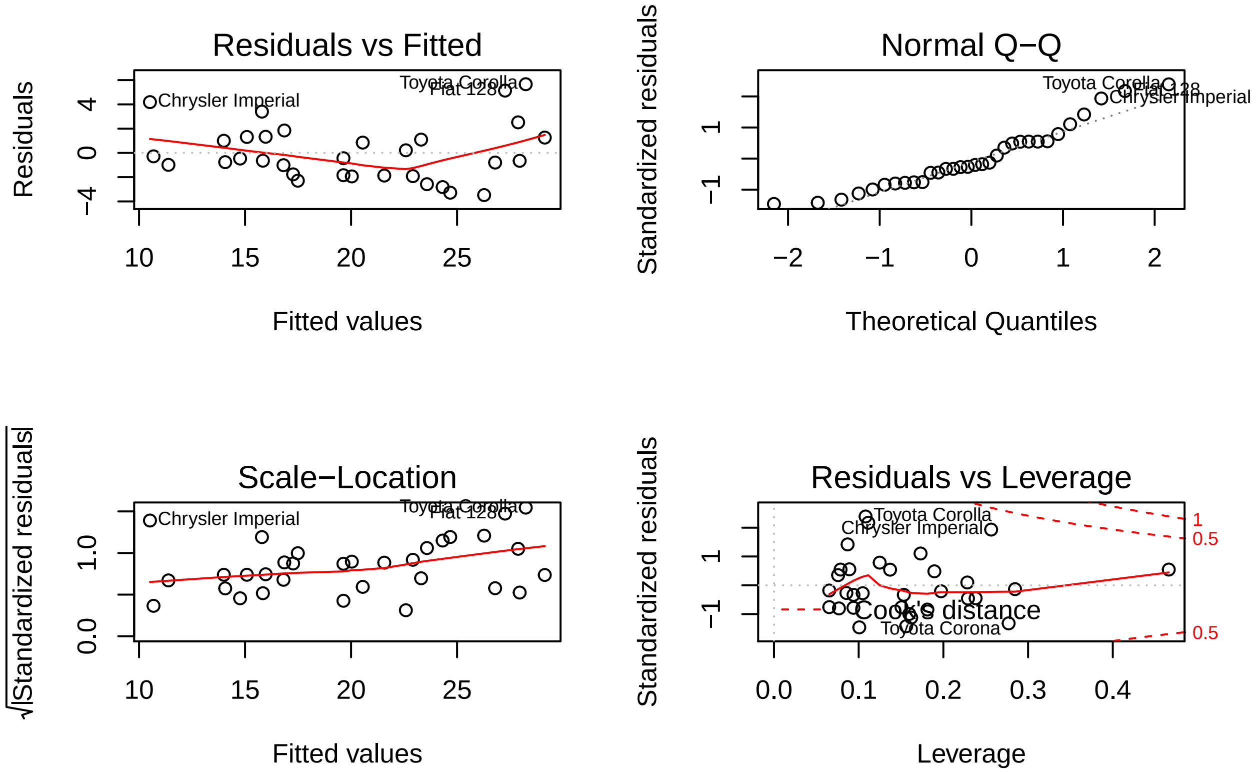
Plot of residuals of single variable linear model

```
ggplot(linear_model) +
        geom_point(aes(x=.fitted, y=.resid)) +
        theme_minimal() +
        labs(title = "Plot of residuals of linear model",
             y = "residuals",
             x = "miles per gallon") +
        scale_colour_continuous(name = "transmission") +
        geom_abline(slope = 0,
                    intercept = 0)
```

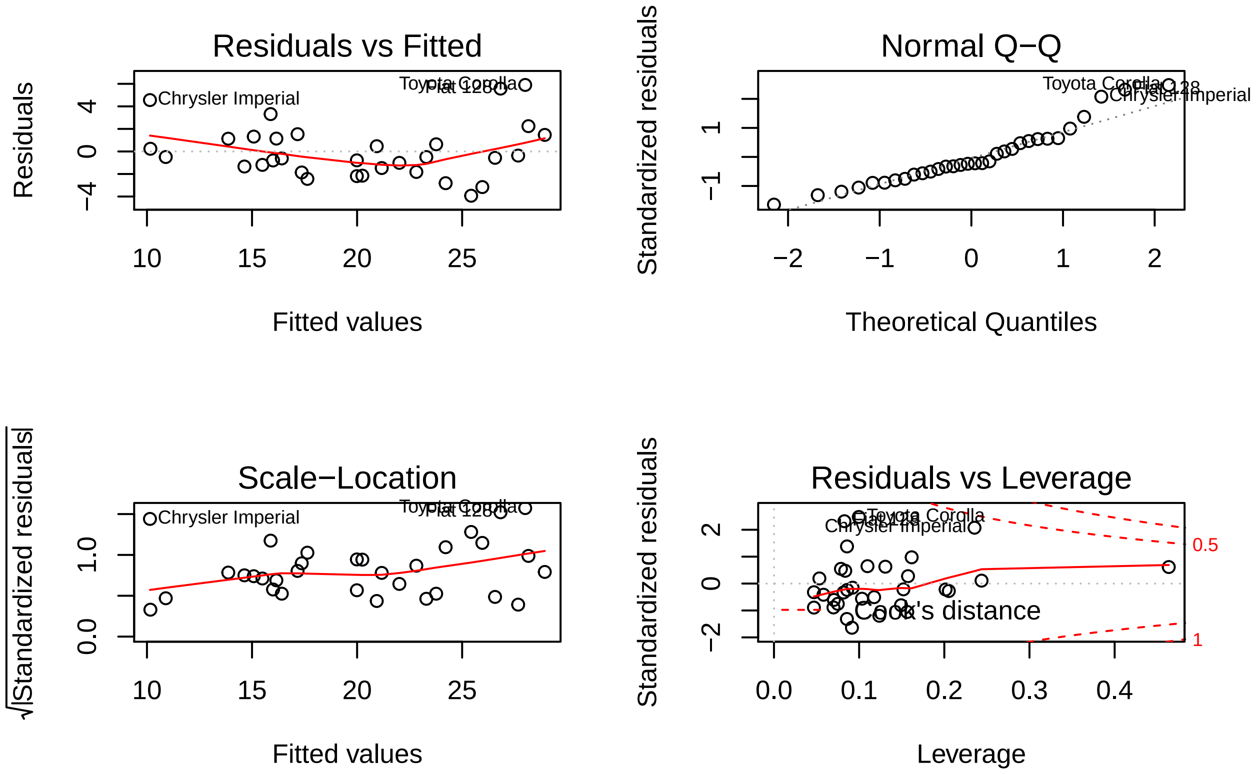


Diagnostic plots for multivariate models 1,2,3,4

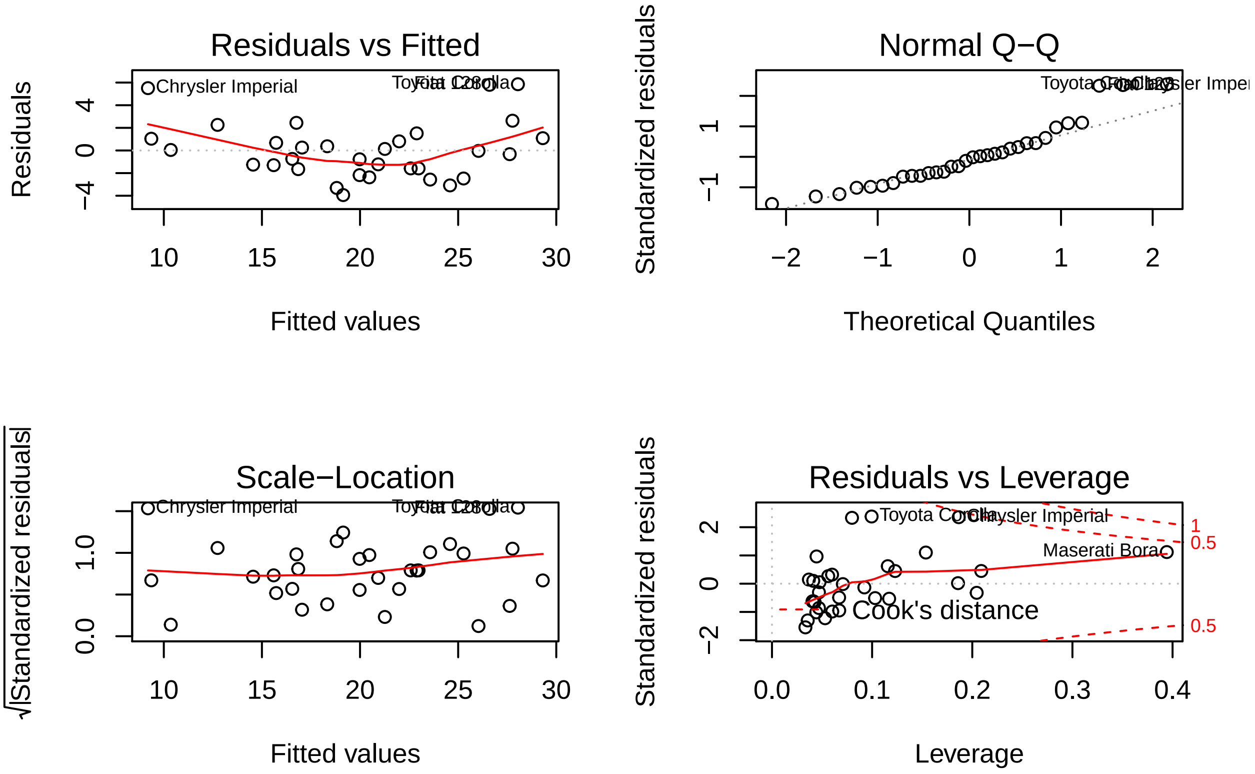
```
par(mfrow = c(2,2))
plot(model_1)
```



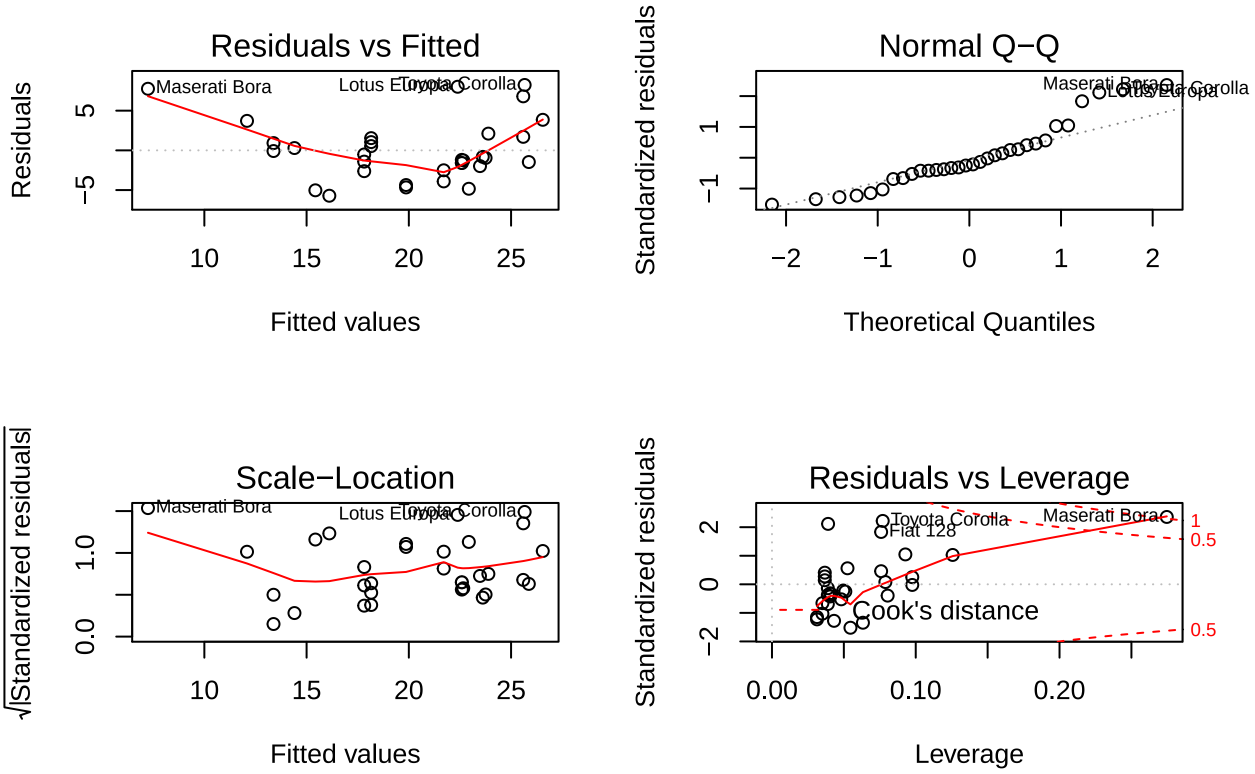
```
plot(model_2)
```



```
plot(model_3)
```


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
−4
0
4
10
15
20
25
30
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Fiat 128
Chrysler Imperial
−1
1
−2
−1
0
1
2
Scale−Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
0.0
1.0
10
15
20
25
30
Residuals vs Leverage
Standardized residuals
Leverage
Toyota Corolla
Chrysler Imperial
Maserati Bora
Cook's distance
1
0.5
0.5
−2
0
2
0.0
0.1
0.2
0.3
0.4

```
plot(model_4)
```


Residuals vs Fitted
Residuals
Fitted values
Maserati Bora
Lotus Europa
Toyota Corolla
−5
5
10
15
20
25
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Maserati Bora
Toyota Corolla
Lotus Europa
−1
1
−2
−1
0
1
2
Scale−Location
√|Standardized residuals|
Fitted values
Maserati Bora
Lotus Europa
Toyota Corolla
0.0
1.0
10
15
20
25
Residuals vs Leverage
Standardized residuals
Leverage
Toyota Corolla
Fiat 128
Maserati Bora
Cook's distance
1
0.5
0.5
−2
0
2
0.00
0.10
0.20