Regression model course project

Stephanie

February 2016

Contents

Executive summary

In this report, we will analyze a data set of a collection of cars. 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. We will explore the relationship between a set of variables and miles per gallon (MPG - outcome) and answer the following two questions:

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

Using some exploratory data and simple linear regression analysis, we will determine that:

- There is a signficant difference between the mean MPG for automatic and manual transmission cars
- Manual transmissions achieve a higher value of MPG compared to automatic transmission.

Loading the mtcars data

```
data(mtcars)
```

Is an automatic or manual transmission better for MPG?

We first look at the data to see if we need to process it.

```
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
```

```
str(mtcars)
```

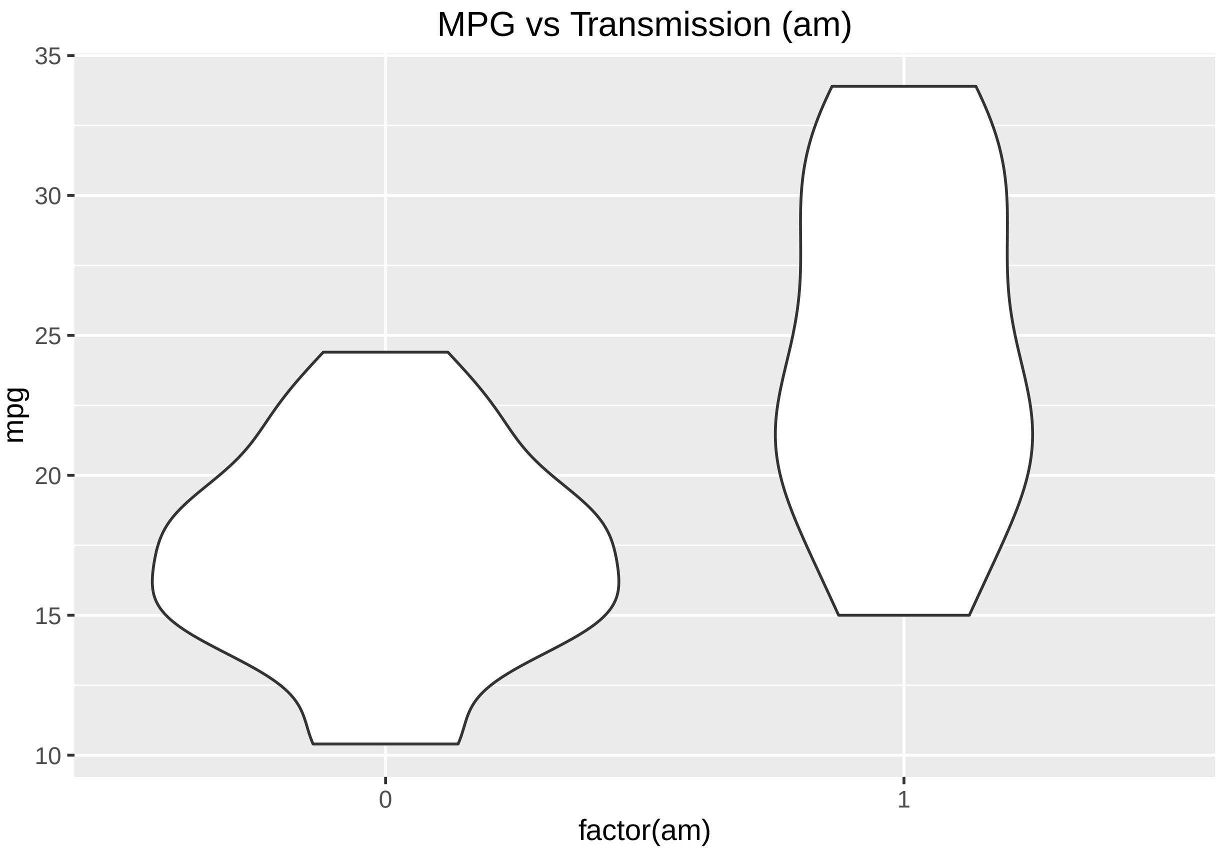
```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 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 ...
```

The data looks good enough to use linear regression. All variables are numerical. Though, we will need to convert some variables as factors along the analysis.

Exploratory data analysis

Let's now look at the relation between MPG (mpg) and transmission (am) with a plot.

```r
library(ggplot2)
plot1 <- ggplot(mtcars, aes(x = factor(am), y = mpg)) + geom_violin() + labs(title = "MPG vs Transmissi
plot1
```



Looking at the plot, we can suppose that manual transmission (am = 1) is associated with a larger MPG.

Simple linear model

Let's fit a simple linear regression model to check this assumption.

```r
lin1 <- lm(mpg ~ factor(am), data=mtcars)
summary(lin1)
```

```
##
## Call:
## lm(formula = mpg ~ factor(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 ***
## factor(am)1    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
```

Let's understand the model:

- if the transmission is automatic (am = 0), the prediction is 17.147, which is the mean of MPG for am = 0
- if the transmission is manual (am = 1), then the prediction is 17.147 + 7.245 = 24.392 which is the mean of MPG for am = 1

Confidence interval

Let's calculate a 95% confidence interval for Beta1.

```r
pe <- coef(summary(lin1))[2,1]
se <- coef(summary(lin1))[2,2]
pe + c(-1,1)*qt(.975,30)*se # (n = 32 ; n-2 = 30)
```

```
## [1]  3.64151 10.84837
```

Conclusion

The p-value (2e10-4 < 0.05) for beta1 is small and the confidence interval does not include zero, so we can reject the null hypothesis in favor of the alternative hypothesis. We therefore assume that there is a significant difference in MPG between the two groups at alpha = 0.05.

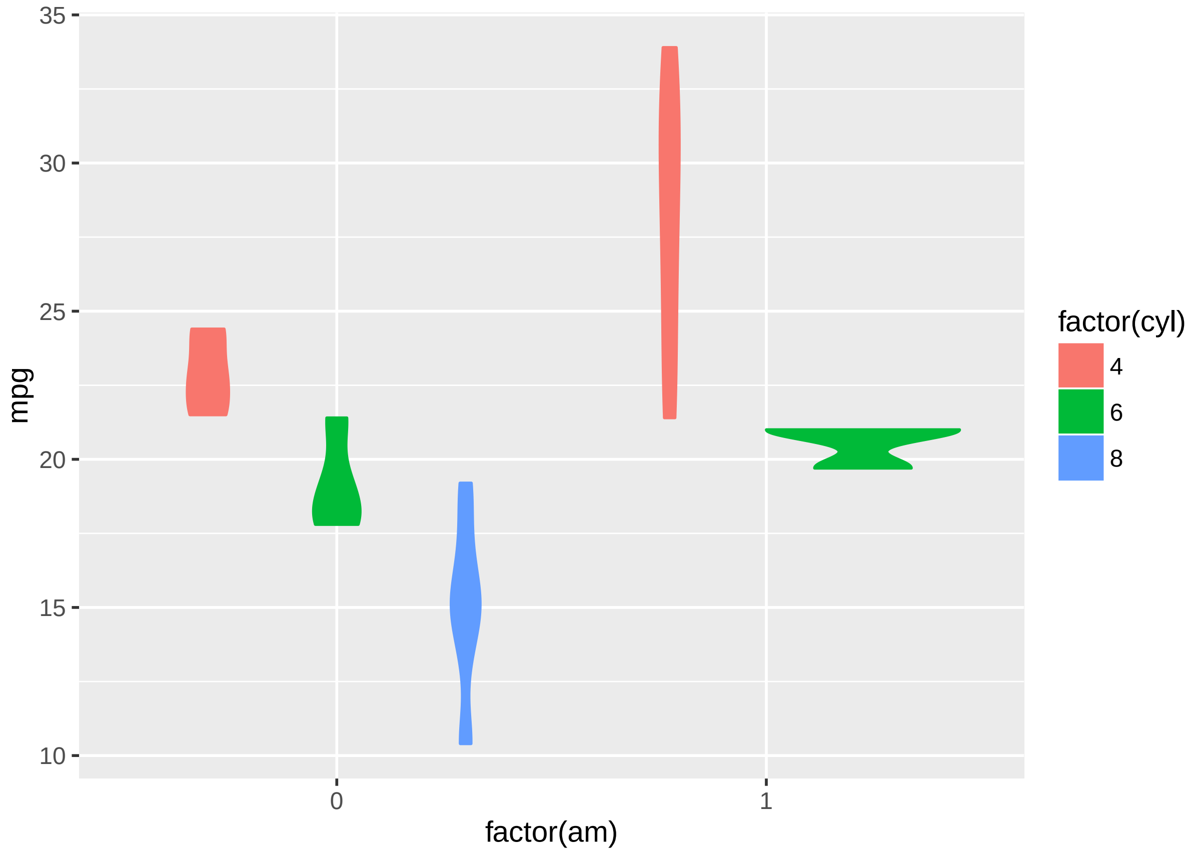
Quantify the MPG difference between automatic and manual transmissions

We have seen that a manual transmission seems to lead to a larger MPG. However, this was a simple model and we need to explore further.

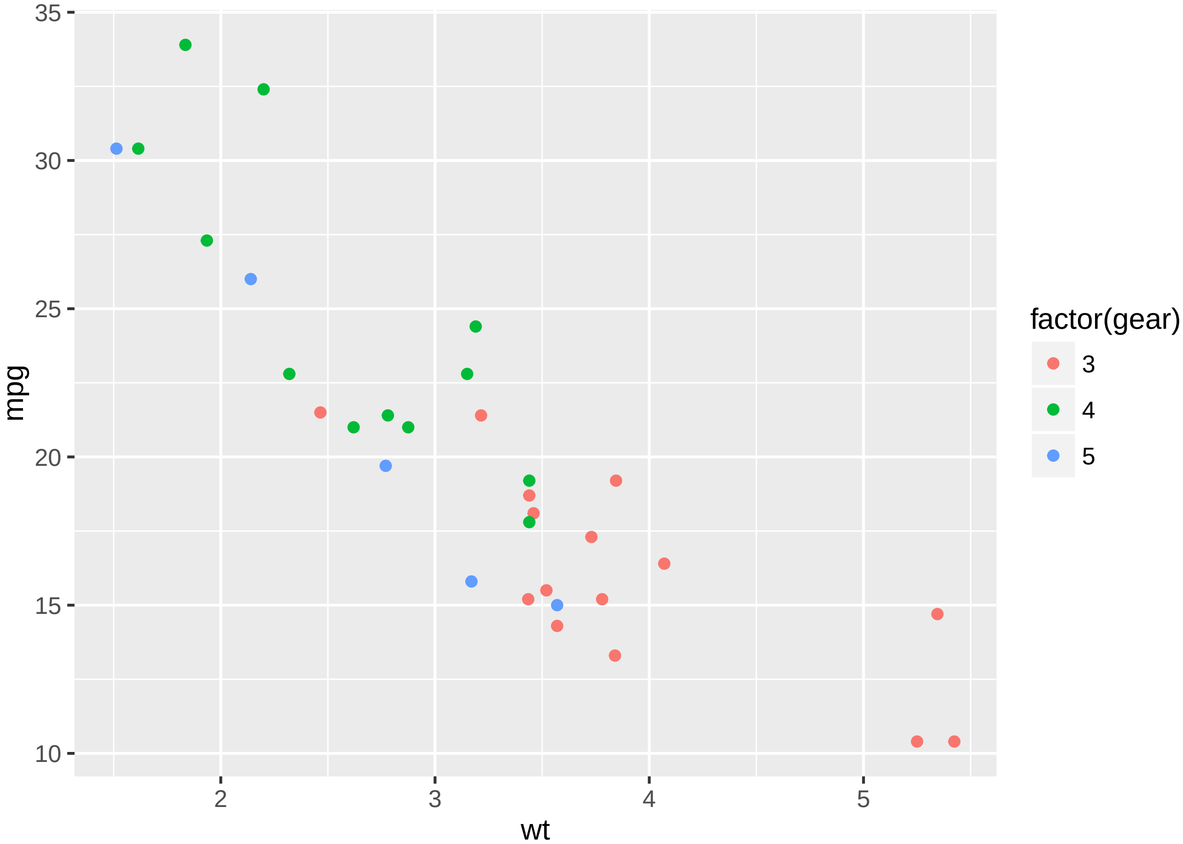
Exploratory data analysis

Let's have a look at the relationships between MPG and other features.

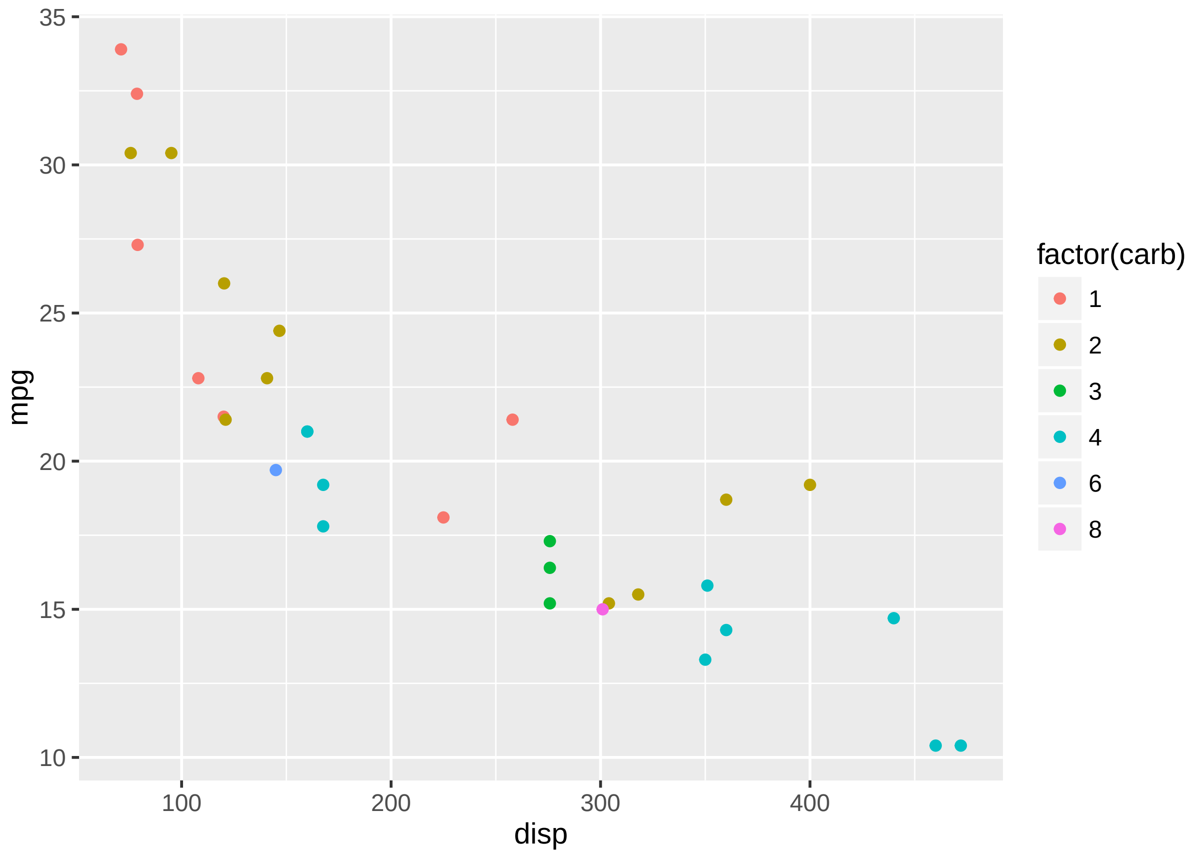
```
plot2 <- ggplot(mtcars, aes(x = factor(am), y = mpg, col = factor(cyl))) + geom_violin(aes(fill = facto
plot2
```




```
plot3 <- ggplot(mtcars, aes(x = wt, y = mpg, col = factor(gear))) + geom_point()
plot3
```




```
plot4 <- ggplot(mtcars, aes(x = disp, y = mpg, col = factor(carb))) + geom_point()
plot4
```



Linear regression models for further exploration

We will create several models and compare them to each other. First, let's see the correlation between the features to choose which one to include into our models.

```r
cor(mtcars)
```

```
##             mpg        cyl       disp         hp        drat         wt
## mpg   1.0000000 -0.8521620 -0.8475514 -0.7761684  0.68117191 -0.8676594
## cyl  -0.8521620  1.0000000  0.9020329  0.8324475 -0.69993811  0.7824958
## disp -0.8475514  0.9020329  1.0000000  0.7909486 -0.71021393  0.8879799
## hp   -0.7761684  0.8324475  0.7909486  1.0000000 -0.44875912  0.6587479
## drat  0.6811719 -0.6999381 -0.7102139 -0.4487591  1.00000000 -0.7124406
## wt   -0.8676594  0.7824958  0.8879799  0.6587479 -0.71244065  1.0000000
## qsec  0.4186840 -0.5912421 -0.4336979 -0.7082234  0.09120476 -0.1747159
## vs    0.6640389 -0.8108118 -0.7104159 -0.7230967  0.44027846 -0.5549157
## am    0.5998324 -0.5226070 -0.5912270 -0.2432043  0.71271113 -0.6924953
## gear  0.4802848 -0.4926866 -0.5555692 -0.1257043  0.69961013 -0.5832870
## carb -0.5509251  0.5269883  0.3949769  0.7498125 -0.09078980  0.4276059
##             qsec         vs          am       gear        carb
## mpg   0.41868403  0.6640389  0.59983243  0.4802848 -0.55092507
## cyl  -0.59124207 -0.8108118 -0.52260705 -0.4926866  0.52698829
## disp -0.43369788 -0.7104159 -0.59122704 -0.5555692  0.39497686
## hp   -0.70822339 -0.7230967 -0.24320426 -0.1257043  0.74981247
## drat  0.09120476  0.4402785  0.71271113  0.6996101 -0.09078980
## wt   -0.17471588 -0.5549157 -0.69249526 -0.5832870  0.42760594
## qsec  1.00000000  0.7445354 -0.22986086 -0.2126822 -0.65624923
```

```
## vs    0.74453544  1.0000000  0.16834512  0.2060233 -0.56960714
## am   -0.22986086  0.1683451  1.00000000  0.7940588  0.05753435
## gear -0.21268223  0.2060233  0.79405876  1.0000000  0.27407284
## carb -0.65624923 -0.5696071  0.05753435  0.2740728  1.00000000
```

The variable the more correlated with MPG is "wt". So we will also add the variable the less correlated with "wt" which is "qsec". And then we add the variable of interest here which is am.

Let's build the models.

```r
lin2 <- lm(mpg ~ wt, data=mtcars)
lin3 <- lm(mpg ~ wt + qsec, data=mtcars)
lin4 <- lm(mpg ~ wt + qsec + factor(am), data=mtcars)

anova(lin2, lin3, lin4)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt
## Model 2: mpg ~ wt + qsec
## Model 3: mpg ~ wt + qsec + factor(am)
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 278.32
## 2     29 195.46  1    82.858 13.7048 0.0009286 ***
## 3     28 169.29  1    26.178  4.3298 0.0467155 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The second (lin3) and third (lin4) models have a p_value very small, which leads us to reject the null hypothesis, and to suppose that these models lead to an improvement in comparison of the model 1.

Let's explore the third model (lin4):

```r
summary(lin4)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + factor(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 ***
## factor(am)1   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
```

All the variables are significant in this model.

Confidence interval

Let's calculate a 95% confidence interval for Beta3.

```r
m <- coef(summary(lin4))[4,1]
se <- coef(summary(lin4))[4,2]
m + c(-1,1)*qt(.975,30)*se # (n = 32 ; n-2 = 30)
```

```
## [1] 0.05438576 5.81728862
```

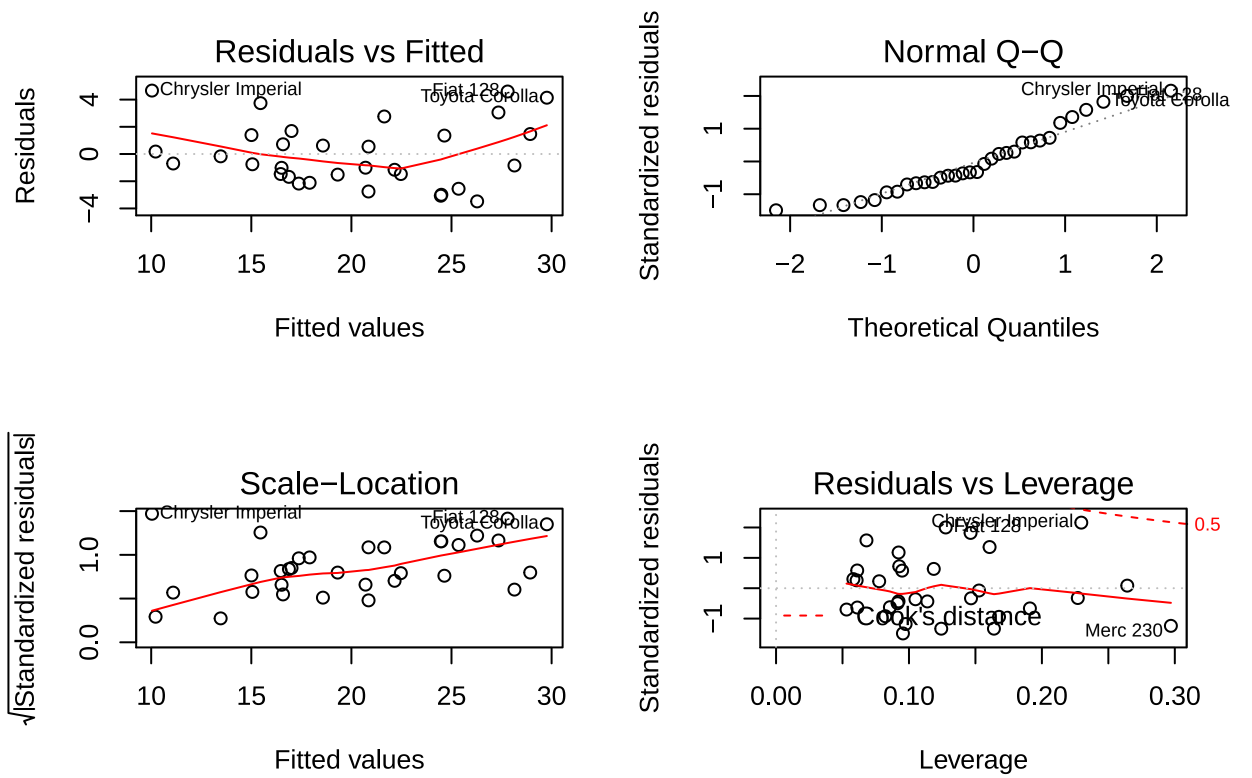
Conclusion

The p-value (4e10-2 < 0.05) for beta1 is small and the confidence interval does not include zero, so we can reject the null hypothesis in favor of the alternative hypothesis. We therefore assume that the means of the 2 groups are significantly different at alpha = 0.05.

Residual plot

Let's plot this model (lin4) to see the residuals.

```r
par(mfrow=c(2,2))
plot(lin4)
```



- *The Residual vs Fitted* and *Scale-Location* plots: there is a slight curve, indicating a slight pattern (to be investigated). And several points seem to be outliers, exercing an influence over the curve, for example Toyota Corolla (row 20) - *Normal Q-Q* plot: the residuals tend to follow a normal distribution, so that the points lie on the line, except for the outliers at the top-right. - *Residuals vs Leverage*: points out those outliers, but indicates that those outliers are within the confidence interval (so not really outliers).