Exploration of the relationship between vehicle's mpg and vehicle's aspects and parameters

Gavriil Shchedrin

11/07/2021

Statement of the problem

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:

- Q1: Is an automatic or manual transmission better for MPG?
- Q2: Quantify the MPG difference between automatic and manual transmissions

Summary of the final model

Lets define the variables:

$$
\begin{aligned}
x_1 &= qsec \\
x_2 &= wt \\
x_3 &= am1 \\
x_4 &= wt * am1
\end{aligned}
$$

Therefore we are working with the model that can be summarized as

$$E[mpg] = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \beta_4 * (x_2 * x_3)$$

For the cars with an automatic transmission we have `x_{3} = 0`:

$$E[mpg|automatic] = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 \simeq 9.72 + 1.01 * x_1 - 2.93 * x_2$$

while the cars with a manual transmission we have `x_{3} = 1`:

$$
\begin{aligned}
E[mpg|manual] &= (\beta_0 + \beta_3) + \beta_1 * x_1 + (\beta_2 + \beta_4) * (x_2) \\
&\simeq (9.72 + 14.07) + 1.01 * x_1 + (-2.93 - 4.14) * x_2
\end{aligned}
$$

Q1: Is an automatic or manual transmission better for MPG?

Answer to Q1: The simple model that fits mpg to the vehicle transmission has a high value of the AIC factor and therefore a better model includes `am`, `qsec`, `wt`, along with the interaction term `wt*am`. Therfore the answer to this question depends on the all of these combined.

Q2: Quantify the MPG difference between automatic and manual transmissions

Answer to Q2: So even though the slope for `E[mpg|manual]` is higher compared to `E[mpg|automatic]`, the high value of intercept indicates that in this dataset the increase of weight of the vehicle results in higher mpg for manual cars. Numerically, the increase in the vehicle wright 1000lbs with an automatic transmission decreases mpg from `9.72` to `6.79` mpg, while the corresponding increase in weight of a vehicle with a manual transmission results in a decrease of the vehicle mpg from `23.79` to `16.72`.

Below we provide 7 key steps that led to this final conclusion.

Step 1: Basic data exploration

```
#loading the libraries

library(UsingR)
library(ggplot2)
library(GGally)
library(dplyr)

#loading the dataset
data(mtcars)
```

Here is the head of the data that we will be woring with

```
head(mtcars,10)
```

```
##                    mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
## Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 230          22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
## Merc 280          19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
```

In total we have 32 different vehilcles and 11 characteristics:

```
dim(mtcars)
```

```
## [1] 32 11
```

We will be focusin on finding a possible relationshuip between these two factors

- [, 1] mpg Miles/(US) gallon
- [, 9] am Transmission (0 = automatic, 1 = manual)

and its possible relationship with the other 9 factors listed in the dataset:

- [, 2] cyl Number of cylinders
- [, 3] disp Displacement (cu.in.)
- [, 4] hp Gross horsepower
- [, 5] drat Rear axle ratio

- [, 6] wt Weight (1000 lbs)
- [, 7] qsec 1/4 mile time
- [, 8] vs Engine (0 = V-shaped, 1 = straight)
- [,10] gear Number of forward gears
- [,11] carb Number of carburetors

Step 2: Advanced data exploration

This is some advacned library with some advacned functionality:

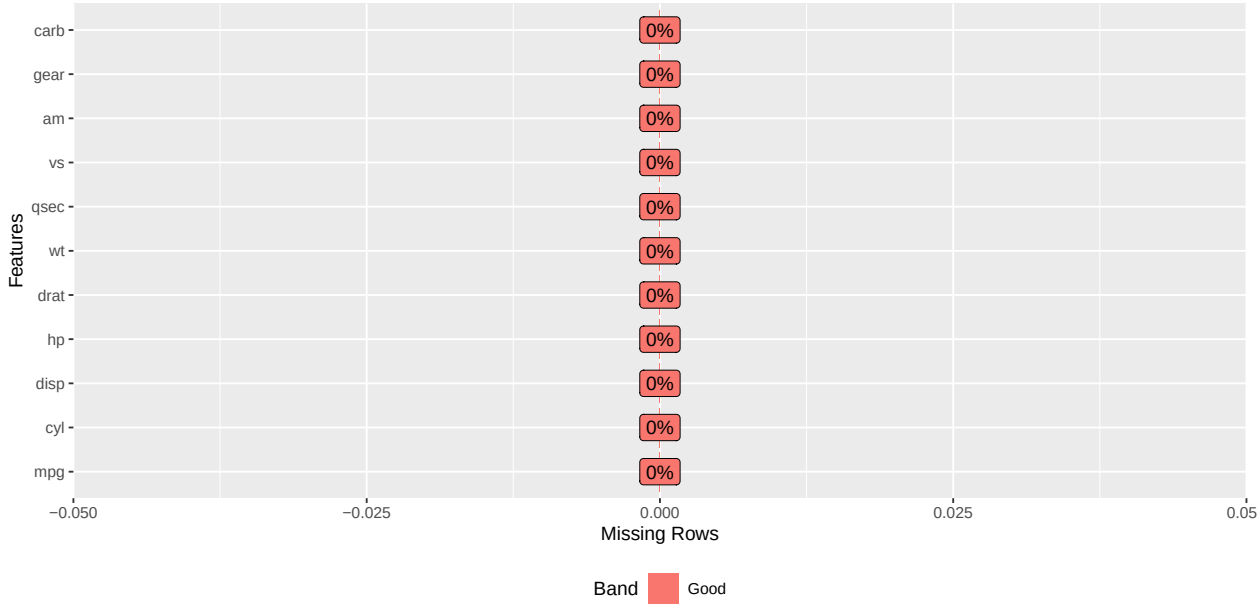
```
# loading the cool EDA library
library(DataExplorer)
```

Here is some cool visual inspection that vaidates our basic EDA:

```
DataExplorer::plot_str(mtcars)
```

Now we shall explore if we have any missing data:

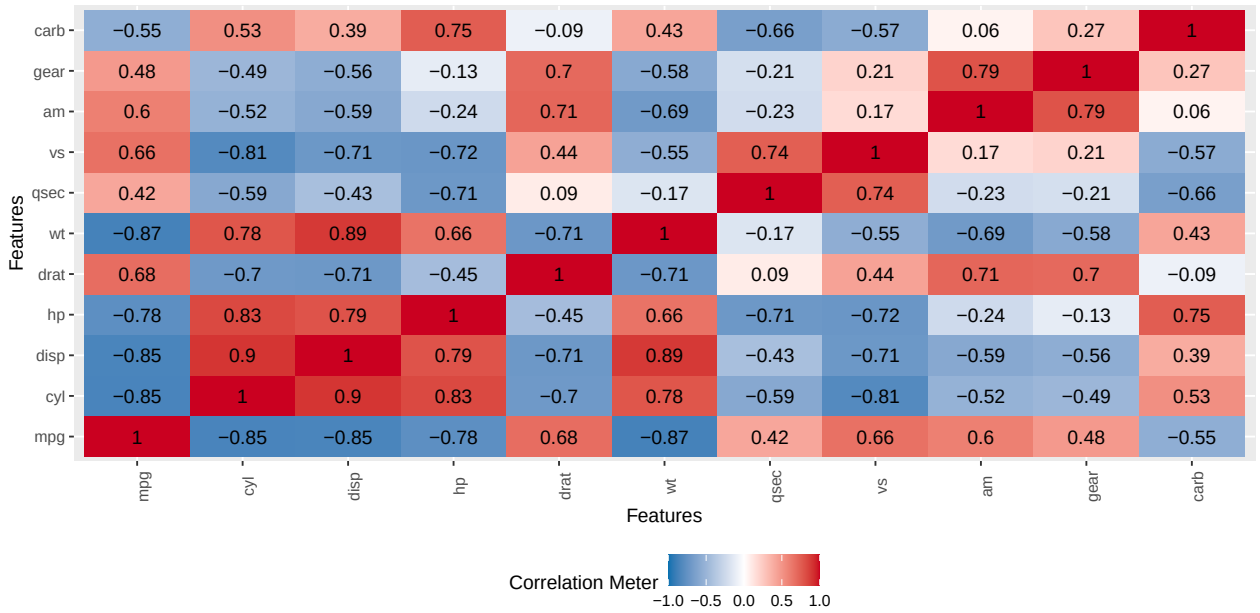
```
DataExplorer::plot_missing(mtcars)
```



Out of all `32*11=352` data points we are not missing any data points!

Lets look at correlations

```
DataExplorer::plot_correlation(mtcars, type=c("all"))
```



Step 3: Turning all the discrete parameters into factor variables

```
# here is the list of discrete variables that should be treated as factors

discreteVariableList <- c("am", "cyl", "gear", "vs")
```

First we need to turn all the discrete variables into factor variables:

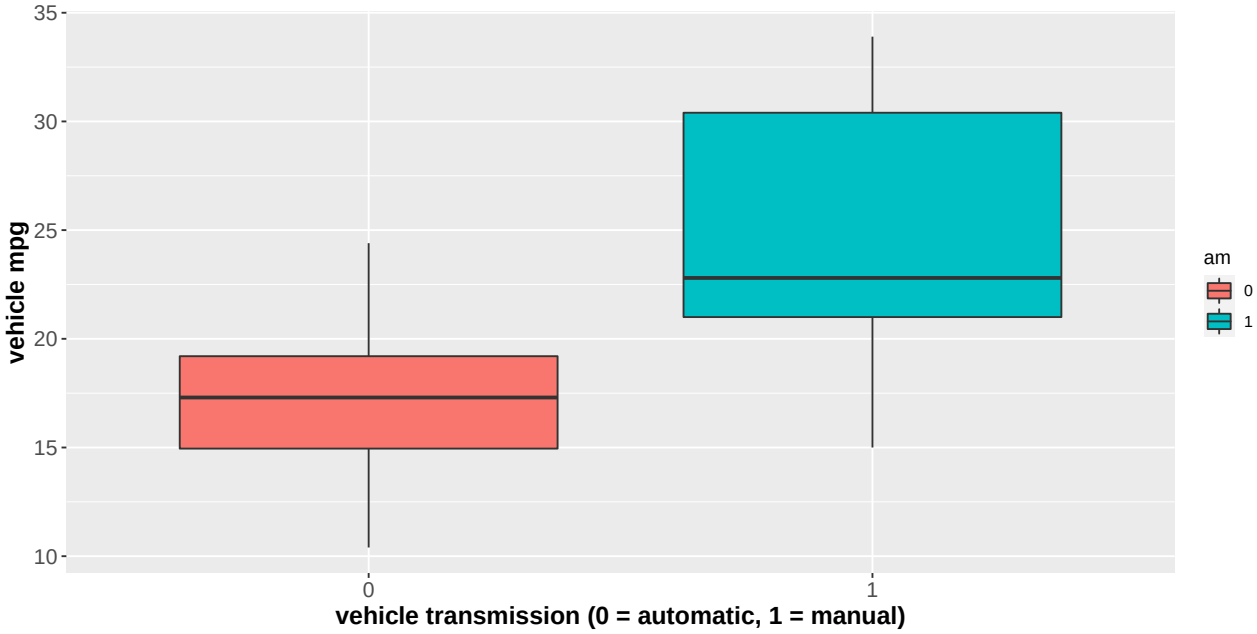
```
mtcarsWithFactors <- mtcars

for(var in discreteVariableList) {
mtcarsWithFactors[,c(var)] <- as.factor(mtcars[,c(var)])
}

head(mtcarsWithFactors)
```

```
##                   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
```

Next we need to plot vehicle mpg vs vs vehicle transmission while completely ignoring all the other parameters:

```
library(ggplot2)
g <- ggplot(mtcarsWithFactors, aes(am, mpg))
g <- g +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=14,face="bold"))
g <- g +
  geom_boxplot(aes(fill = am)) +
  xlab("vehicle transmission (0 = automatic, 1 = manual)") +
  ylab("vehicle mpg")
```

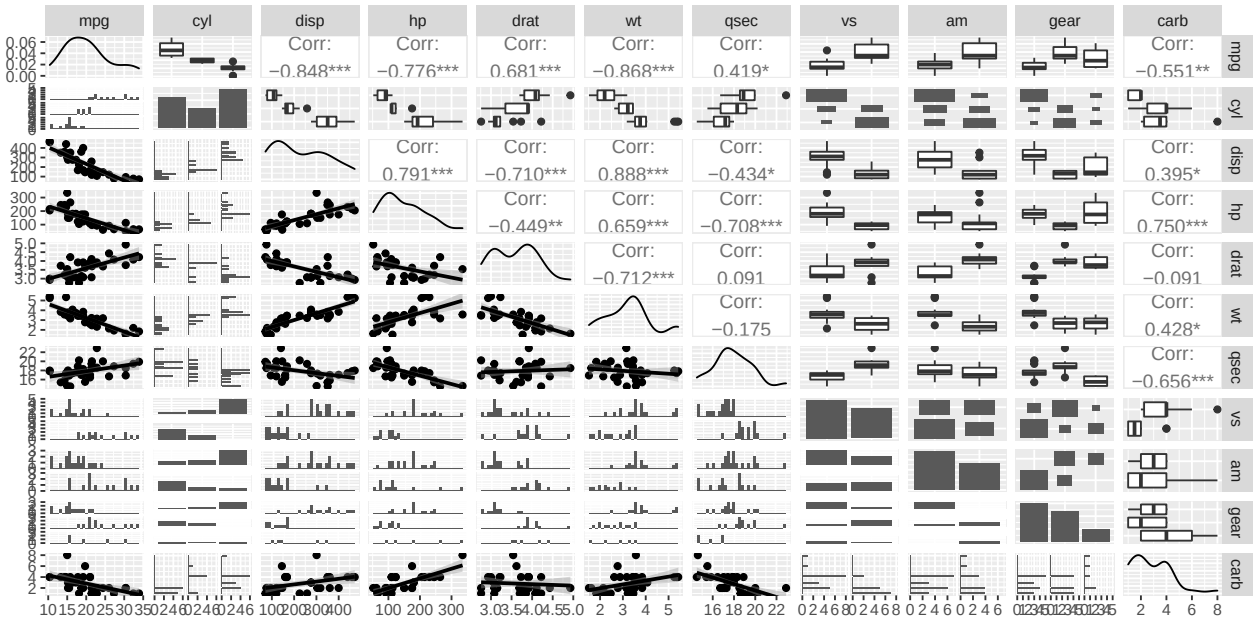
```
g
```



Step 4: The multivariate analysis

Here is the basic ggpairs plot without differentiation by the `am` factor

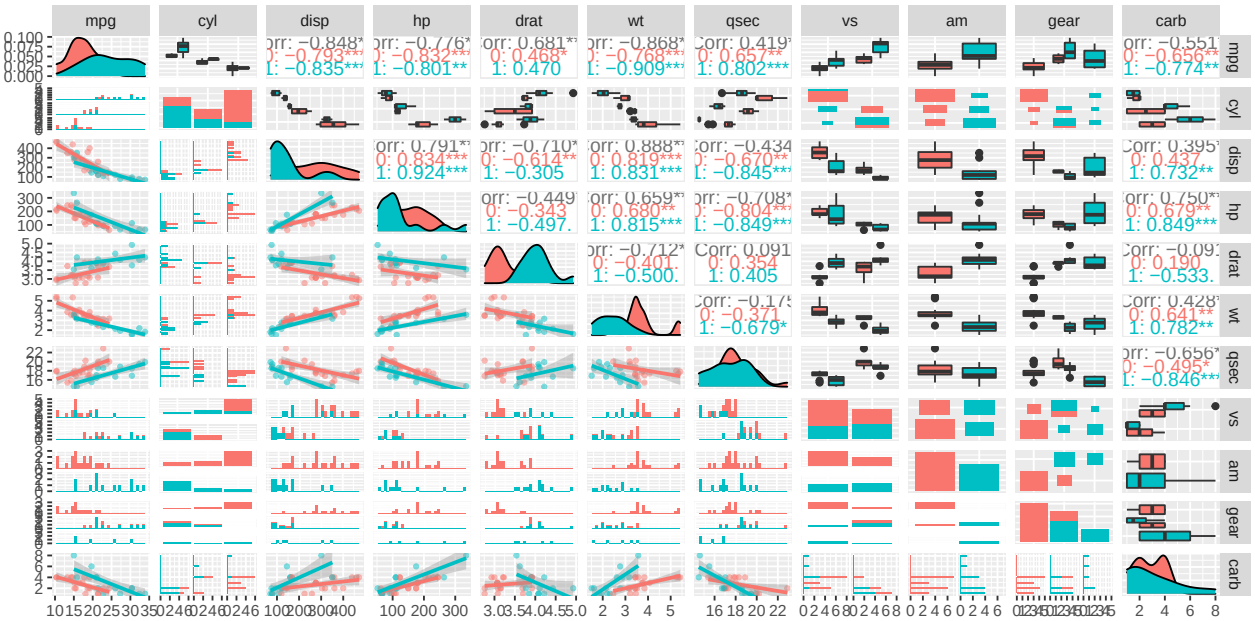
```
ggpairs(mtcarsWithFactors, lower = list(continuous = "smooth"))
```



and here is the ggpairs plot with differentiation

```
ggpairs(data = mtcarsWithFactors,
     mapping = ggplot2::aes(color = am),
     lower = list(continuous = wrap("smooth", alpha=0.5, size=1))
```

```
)
```



Step 5A: Linear model building without any interaction terms

Here is the most basic model that includes `am` as the only (factor) variable:

```
basicModel<-lm(mpg ~ am, data = mtcarsWithFactors)
summary(basicModel)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcarsWithFactors)
##
## 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 ***
## am1            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
```

First of all the p-value of this model is:

$$p_{model} = 0.000285 < \alpha_0 = 0.001$$

which means that with the confidence level 99.9% we reject the Null Hypothesis and therefore the transmission type is indeed a significant factor for the vehicle's mpg.

The Adjusted R-squared is only `0.3385` so we shall consider the model that includes all the variables in the data set:

```
completeModel <- lm(mpg ~ ., data = mtcarsWithFactors)
summary(completeModel)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcarsWithFactors)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2015 -1.2319  0.1033  1.1953  4.3085
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.09262   17.13627   0.881   0.3895
## cyl6        -1.19940    2.38736  -0.502   0.6212
## cyl8         3.05492    4.82987   0.633   0.5346
## disp         0.01257    0.01774   0.708   0.4873
## hp          -0.05712    0.03175  -1.799   0.0879 .
## drat         0.73577    1.98461   0.371   0.7149
## wt          -3.54512    1.90895  -1.857   0.0789 .
## qsec         0.76801    0.75222   1.021   0.3201
## vs1          2.48849    2.54015   0.980   0.3396
## am1          3.34736    2.28948   1.462   0.1601
## gear4       -0.99922    2.94658  -0.339   0.7382
## gear5        1.06455    3.02730   0.352   0.7290
## carb         0.78703    1.03599   0.760   0.4568
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.616 on 19 degrees of freedom
## Multiple R-squared:  0.8845, Adjusted R-squared:  0.8116
## F-statistic: 12.13 on 12 and 19 DF,  p-value: 1.764e-06
```

As we can see the Adjusted R-squared of the `completeModel` is `0.8116` which is significantly higher than the Adjusted R-squared of the `basicModel`

The Stepwise Algorithm allows one to choose the optimal set of the variables from the `completeModel` based on the Akaike information criterion (AIC):

$$\text{AIC} = 2k - 2\ln(\hat{L})$$

Here

$$k$$

is the number of estimated parameters in the statistical model and

$$\hat{L}$$

is the maximum value of the likelihood function for the model.

```
aicModelFit <- step(completeModel, trace=FALSE)

summary(aicModelFit)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcarsWithFactors)
##
## 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 ***
## am1           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
```

The Adjusted R-squared of the model that is optimized by the AIC criteria is `0.8336` and in this model the key parameters that affect `mpg` are not only `am` but also `wt` and `qsec`.

Step 5B: Linear model building with interaction between the key variables

Lets build a model with all possible interactions:

```
interactionModel <- lm(formula = mpg ~ wt + qsec + am, data = mtcarsWithFactors)
interactionModelOne <- lm(formula = mpg ~ wt*qsec + am, data = mtcarsWithFactors)
interactionModelTwo <- lm(formula = mpg ~ wt + qsec*am, data = mtcarsWithFactors)
interactionModelThree <- lm(formula = mpg ~ qsec + wt*am, data = mtcarsWithFactors)
```

```
AIC(interactionModelOne)[1] > AIC(interactionModelTwo)[1]
```

```
## [1] TRUE
```

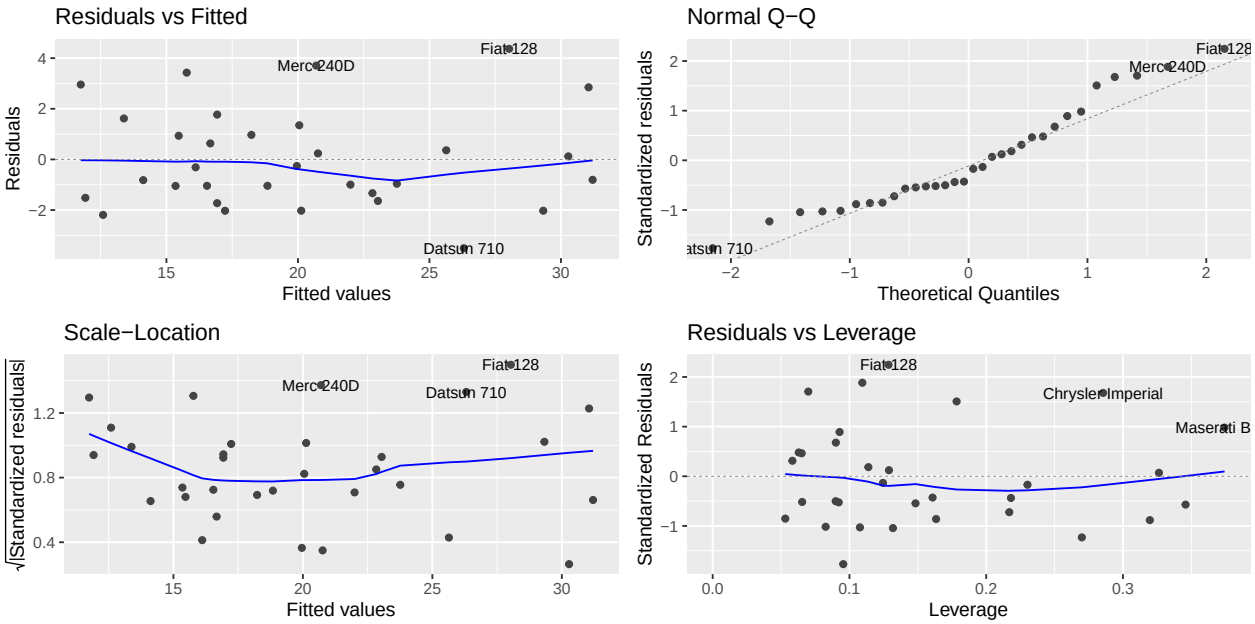
```
AIC(interactionModelTwo)[1] > AIC(interactionModelThree)[1]
```

```
## [1] TRUE
```

Since the Akaike's Information Criterion (AIC) is lowest for the last model `interactionModelThree` which is built on the linear model `formula = mpg ~ qsec + wt*am` we choose this model as the final model for predicting the effect of `qsec`,am,wt, `and the interaction`wt*am‘ on the vehicle mpg.

Step 6: The QQ plot

```
library(ggplot2)
library(ggfortify)

autoplot(interactionModelThree, label.size = 3)
```



Step 7: Interpretation of the final model

```
summary(interactionModelThree)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.723053  5.8990407  1.648243 0.1108925394
## qsec         1.016974  0.2520152  4.035366 0.0004030165
## wt          -2.936531  0.6660253 -4.409038 0.0001488947
## am1         14.079428  3.4352512  4.098515 0.0003408693
## wt:am1      -4.141376  1.1968119 -3.460340 0.0018085763
```

Lets define the variables:

$$x_1 = qsec$$
$$x_2 = wt$$
$$x_3 = am1$$
$$x_4 = wt * am1$$

Therefore we are working with the model that can be summarized as

$$E[mpg] = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \beta_4 * (x_2 * x_3)$$

For the cars with an automatic transmission we have `x_{3} = 0`:

$$E[mpg|automatic] = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 \simeq 9.72 + 1.01 * x_1 - 2.93 * x_2$$

while the cars with a manual transmission we have `x_{3} = 1`:

$$
\begin{aligned}
E[mpg|manual] &= (\beta_0 + \beta_3) + \beta_1 * x_1 + (\beta_2 + \beta_4) * (x_2) \\
&\simeq (9.72 + 14.07) + 1.01 * x_1 + (-2.93 - 4.14) * x_2
\end{aligned}
$$

Q1: Is an automatic or manual transmission better for MPG?

Answer to Q1: The simple model that fits mpg to the vehicle transmission has a high value of the AIC factor and therefore a better model includes `am`, `qsec`, `wt`, along with the interaction term `wt*am`. Therfore the answer to this question depends on the all of these combined.

Q2: Quantify the MPG difference between automatic and manual transmissions

Answer to Q2: So even though the slope for `E[mpg|manual]` is higher compared to `E[mpg|automatic]`, the high value of intercept indicates that in this dataset the increase of weight of the vehicle results in higher mpg for manual cars. Numerically, the increase in the vehicle wright 1000lbs with an automatic transmission decreases mpg from `9.72` to `6.79` mpg, while the corresponding increase in weight of a vehicle with a manual transmission results in a decrease of the vechile mpg from `23.79` to `16.72`.