Regression Models Peer Assessment

Emanuel Calvo

November 21, 2015

Context

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

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

Your report must be:

- Written as a PDF printout of a compiled (using knitr) R markdown document.
- Brief. Roughly the equivalent of 2 pages or less for the main text. Supporting figures in an appendix can be included up to 5 total pages including the 2 for the main report. The appendix can only include figures.
- Include a first paragraph executive summary.

Executive Summary

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). This dataset is included in the `datasets` library.

From the research done in the current document, we are in place to confirm that the manual transmission in average is better for the consupmtion. However, this trend could be reverted with transmissions with 4 gears.

Exploratory data analysis

We can observ that the transmission systems come from different populations, as the p-value over a t-test give us 0.0013736 between Miles per Galon and Transmissions. Also the mean of each group has 7.2449393 of difference.

All the correlations against mpg:

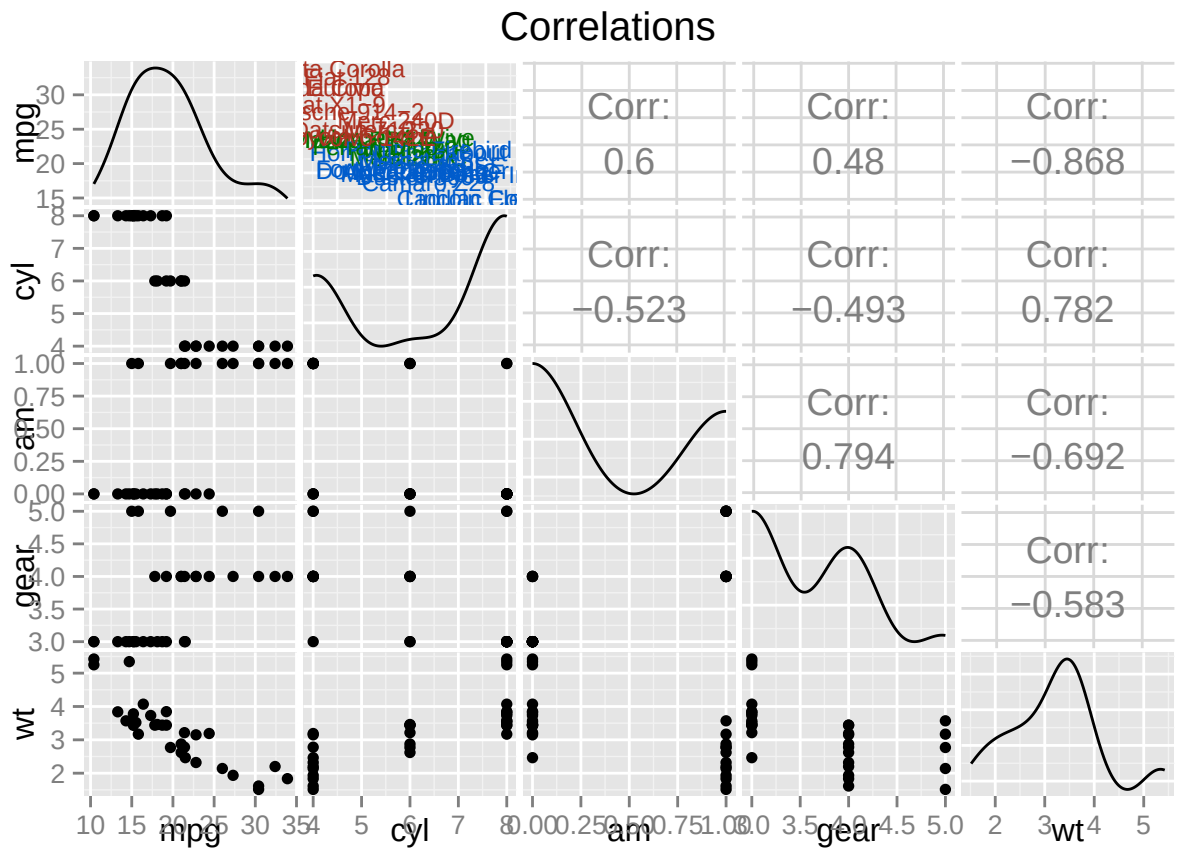
```
cor(mtcars)[1,]
```

```
##        mpg        cyl       disp         hp       drat         wt
##  1.0000000 -0.8521620 -0.8475514 -0.7761684  0.6811719 -0.8676594
##       qsec         vs         am       gear       carb
##  0.4186840  0.6640389  0.5998324  0.4802848 -0.5509251
```

Let's consider the following figure with the correlations among a selected set of columns from the dataset:

```
g <- ggpairs(mtcars[,c("mpg","cyl","am", "gear","wt")], title = "Correlations" )
  plot <- ggplot2::ggplot(mtcars, ggplot2::aes(x=wt, y=mpg, label=rownames(mtcars)))
  plot <- plot +
    ggplot2::geom_text(ggplot2::aes(colour=factor(cyl)), size = 3) +
    ggplot2::scale_colour_discrete(l=40)
g <- putPlot(g, plot, 1, 2)

g
```



Raw model:

```
Model <- lm(mpg ~ ., data=mtcars) # Hidden output.
summary(Model)
```

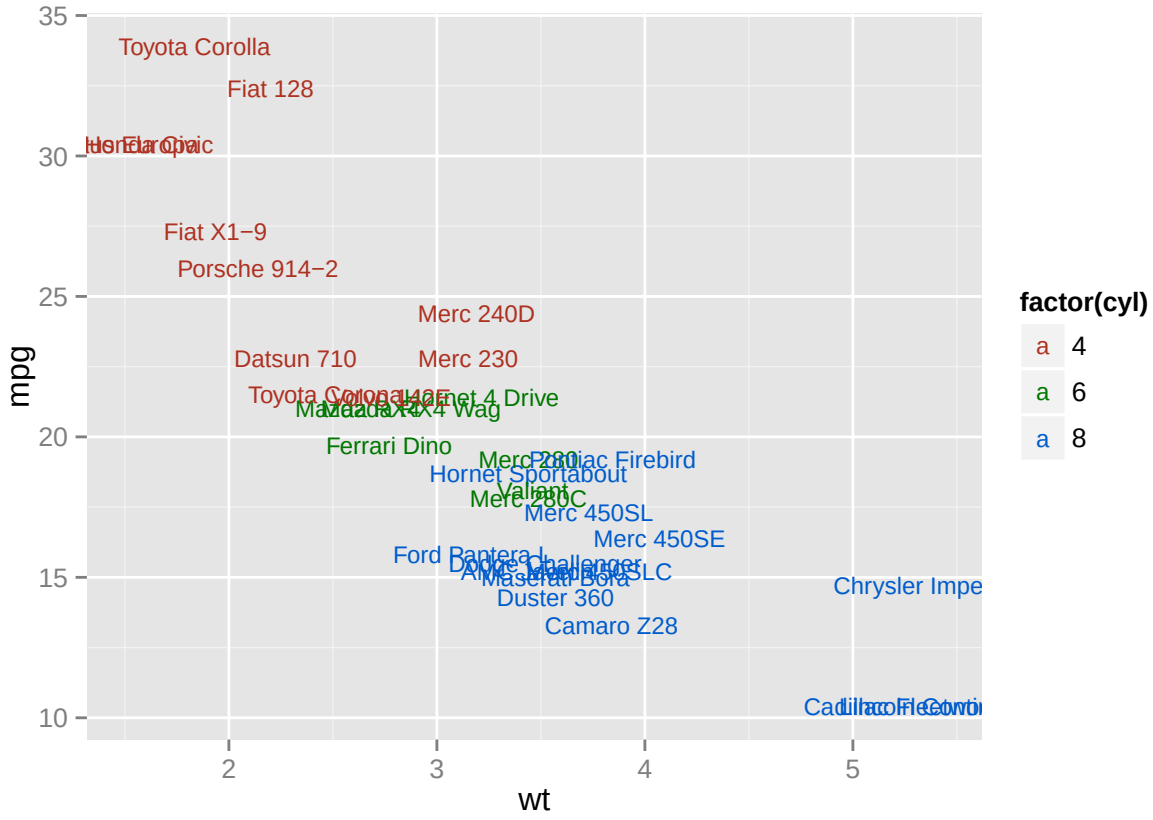
```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
```

```
## 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
## vs           0.31776    2.10451   0.151   0.8814
## am           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
```

Some indicators

There is a clear relationship between the performance over Miles per Galon with the weight of the car. Also, weighter cars needs more cylinders, which clearly affects the consumption per mile.

```
g2 <- ggplot2::ggplot(mtcars, ggplot2::aes(x=wt, y=mpg, label=rownames(mtcars))) +
      ggplot2::geom_text(ggplot2::aes(colour=factor(cyl)), size = 3) +
      ggplot2::scale_colour_discrete(l=40)
g2
```

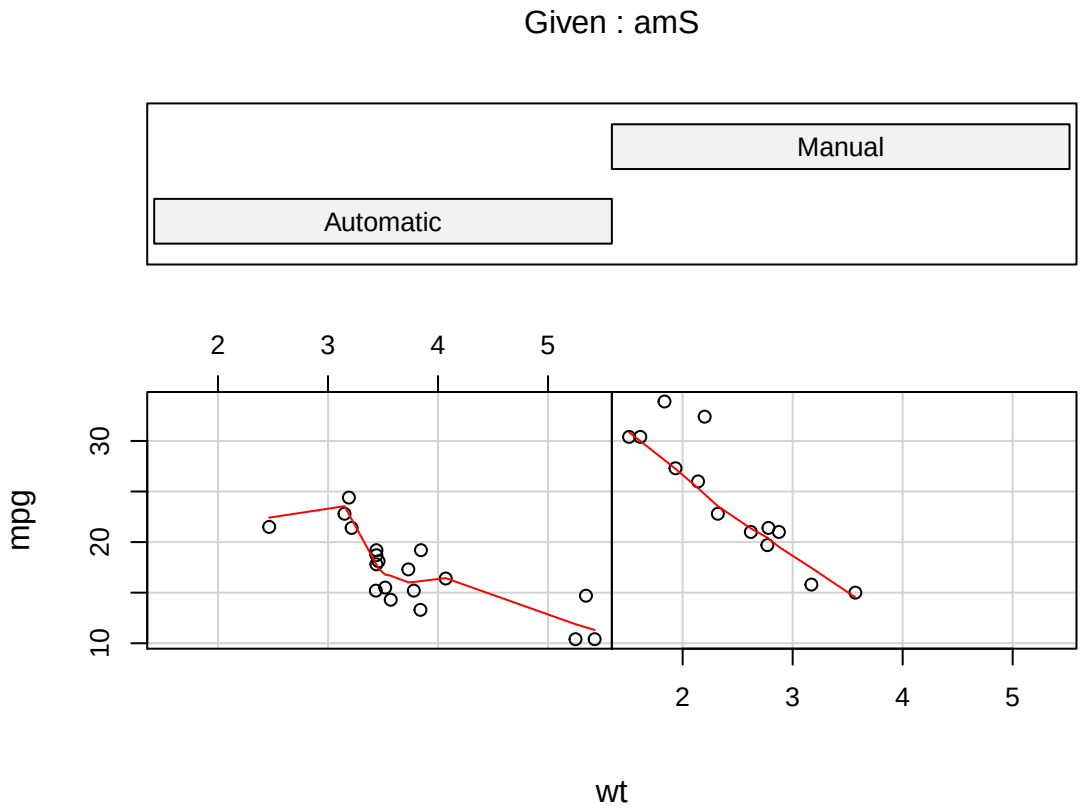


How the transmission system affects the Miles/(US) gallon

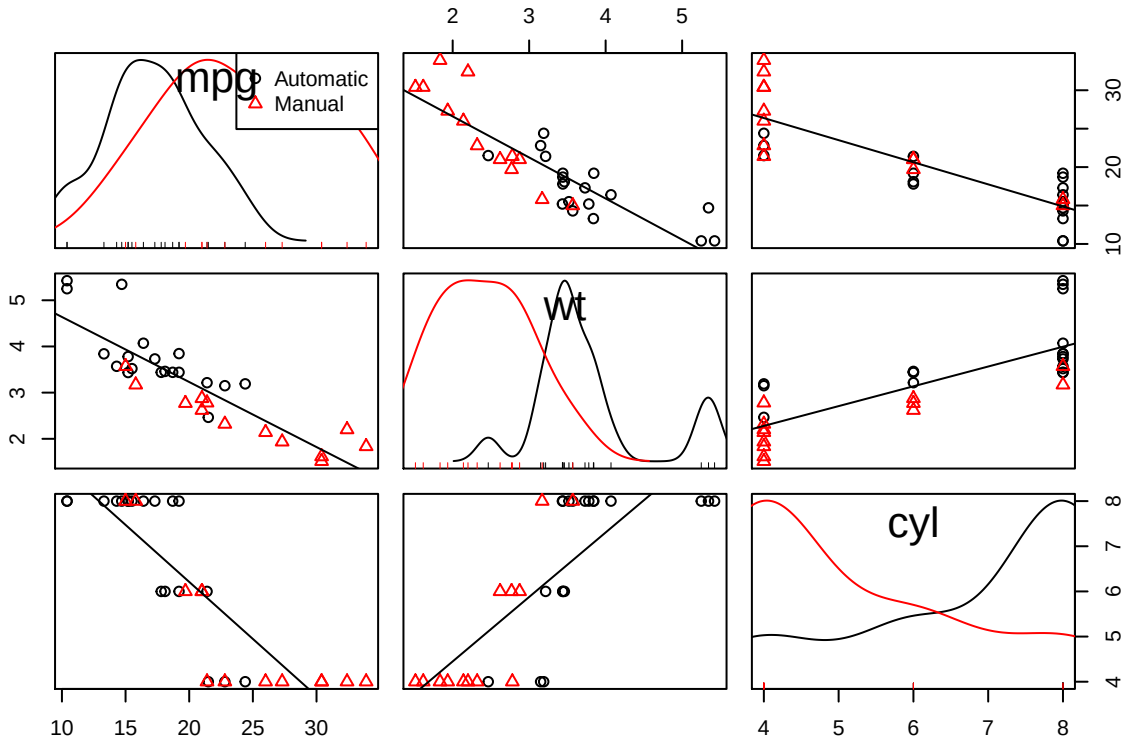
Automatic trassmission with low gears (3) have a bad performance comparing with 4 gears in manual transmission. However, the tendency shows that manual transmissions with higher gears have more variablity and worser performance.

Also, new automatic systems with 4 gears, have a better performance in consume with higher gears.

```
coplot(mpg ~ wt | amS, data = mtcars,
       panel = panel.smooth, rows = 1)
```




```
spm(~ mpg + wt + cyl | amS,data = mtcars,smoother=FALSE)
```

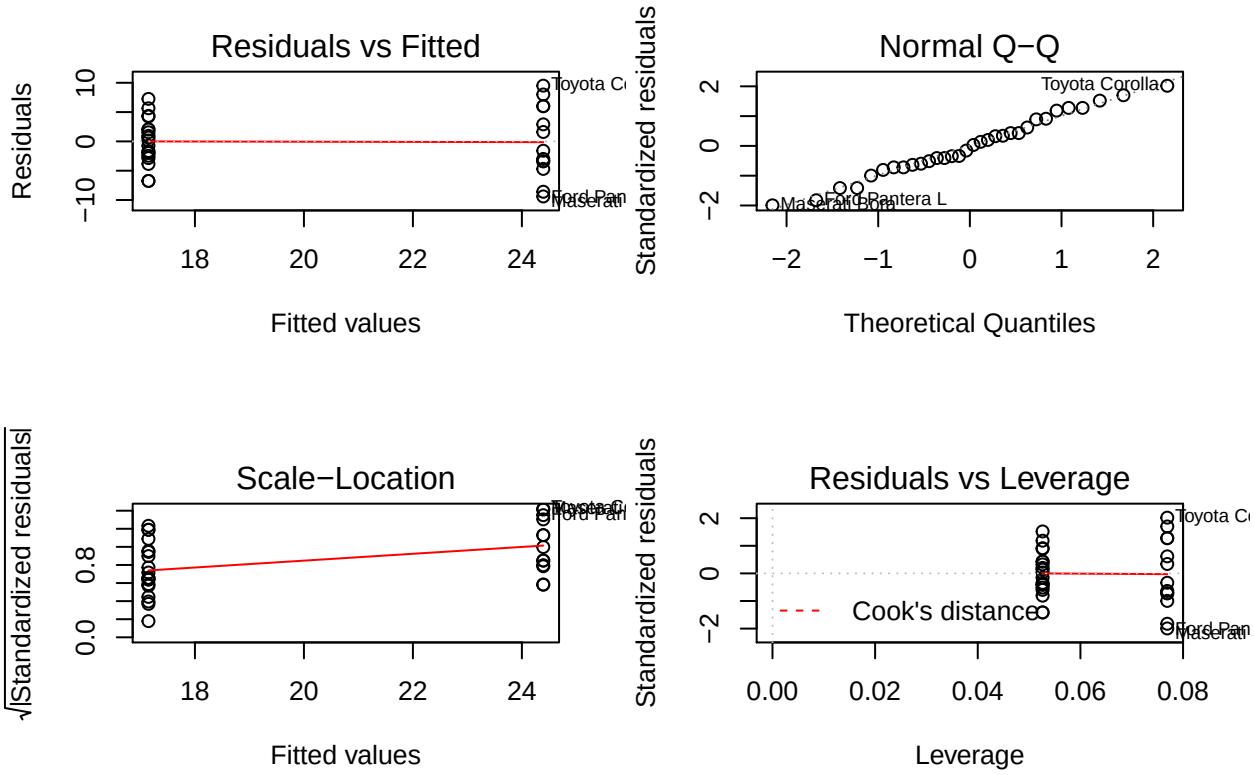


Quantifying the MPG difference between automatic and manual transmissions

As first approach, the following code fits a simple model using the transmission as the predictor and the Miles per galon as the outcome.

```
transModel <- lm(mpg ~ amS, data=mtcars)

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




```
transModelSum <- summary(transModel)
```

On average, a car with automatic transmission has 17.1473684mpg, for manual transmission cars the average is 24 mpg, with an increase of 7.2449393 mpg.

However, the adjusted R-squared value shows that this model explain only the 34 % of the variation on the Miles per Galon.

Found a better model:

```
expModel <- lm(mpg ~ wt +  qsec  + am +  wt:am, data=mtcars)
sumExpModel <- summary(expModel)
sumExpModel
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am + wt:am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5076 -1.3801 -0.5588  1.0630  4.3684
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.723      5.899   1.648 0.110893
## wt            -2.937      0.666  -4.409 0.000149 ***
```

```
## qsec           1.017      0.252   4.035 0.000403 ***
## am            14.079      3.435   4.099 0.000341 ***
## wt:am         -4.141      1.197  -3.460 0.001809 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.084 on 27 degrees of freedom
## Multiple R-squared:  0.8959, Adjusted R-squared:  0.8804
## F-statistic: 58.06 on 4 and 27 DF,  p-value: 7.168e-13
```

This model explains the 88 of the variance of the Miles per Galon.

Model proposed to quantify the difference betwen Miles per galon and transmission

We compare both models (simple linear regression using mpg and transmission system and a custom model described here) using the following functions and code:

```
anova(transModel,expModel)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ amS
## Model 2: mpg ~ wt + qsec + am + wt:am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     27 117.28  3    603.62 46.323 8.847e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

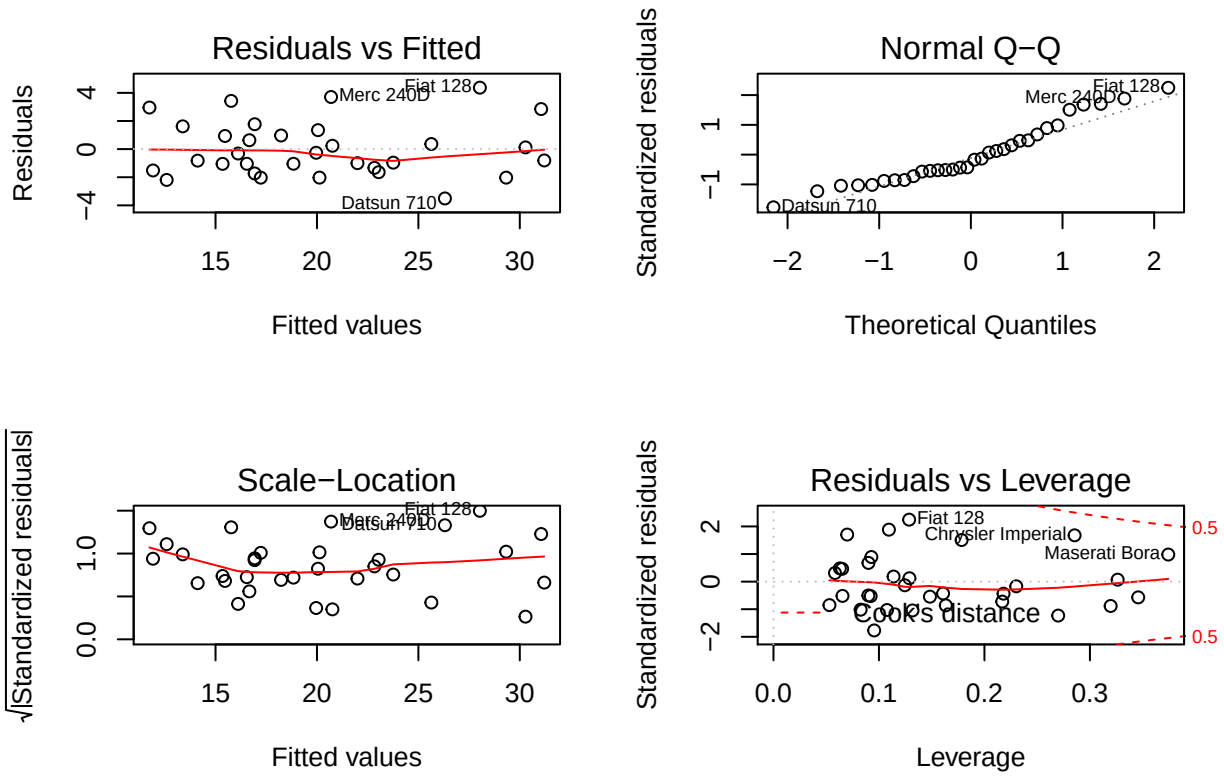
```
anovaModels <- anova(transModel,expModel)
summary(expModel)$coef
```

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

The best fit model we found is within the formula "mpg ~ wt + qsec + ams + wt:am".

Residuals for the selected multivariate model

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



From the graph above and details collected we canmake the following assumptions:

1. Reisudals vs Fitted does not show any visible pattern.
2. The normal Q-Q shows that the resduals are nornally distributed.
3. The Scale-Location confirm that the points are randomly distributed and that there is a constant variance.
4. The Residuals vs Leverage shows that there are no significant outliers on the population, as the values are within the 0.5 range.

As for the DfBetas (measure on effect over an observation by the estimation of a regression coefficient), we obtain 0 which means that our model is a good fit for this population.