Regression Models: Course Project

Amit Kohli

February 15, 2016

Executive Summary This report analyzes mtcars dataset to explore if an automatic or manual transmission is better for cars in terms of miles per gallon (mpg) and quantifying the difference between both types of transmissions. The reports uncovers a strong relationship between vehicle weight and transmission type, but there is still a statistically significant increase in the mean value of gas mileage attributable to manual transmissions – *between 0.05 and 5.8 mpg.*

Exploratory Data Analysis Dataset has 32 observations on 11 variables.

```
## [1] 32 11
##                    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
##                  mpg cyl  disp  hp drat    wt qsec vs am gear carb
## Porsche 914-2  26.0   4 120.3  91 4.43 2.140 16.7  0  1    5    2
## Lotus Europa   30.4   4  95.1 113 3.77 1.513 16.9  1  1    5    2
## Ford Pantera L 15.8   8 351.0 264 4.22 3.170 14.5  0  1    5    4
## Ferrari Dino   19.7   6 145.0 175 3.62 2.770 15.5  0  1    5    6
## Maserati Bora  15.0   8 301.0 335 3.54 3.570 14.6  0  1    5    8
## Volvo 142E     21.4   4 121.0 109 4.11 2.780 18.6  1  1    4    2
```

By checking correlations, weight variable affects the fuel economy the most - and by plotting mpg and weight against transmission types (*Figures 1 & 2*) – the report reveals that weight difference for transmissions is stronger than mpg difference.

Inference The report makes the null hypothesis as the MPG of the automatic and manual transmissions are from the same population by utilizing the the two sample T-test function. Since the p-value is 0.00137, we reject our null hypothesis. So, the automatic and manual transmissions are from different populations. And the mean for MPG of manual transmitted cars is about 7 more than that of automatic transmitted cars.

```
## [1] 0.001373638
## mean of x mean of y
##  17.14737  24.39231
```

Regression Analysis The report runs severals models to understand relationship between different variables, starting with the full model below. As the p-value is much less than 0.05, we reject the null hypothesis. Hence there is a significant relationship between the variables in the linear regression model of the dataset faithful. Results of **Model 1**:

```
## lm(formula = mpg ~ ., data = mtcars)
## [1] "Residual Standard Error:  2.83316868537992"
## [1] "Degrees of Freedom:  15"
## [1] "Adjusted R-squared:  0.779021526312067  (model can explain 78% of the variance of the MPG variable)"
## [1] "p-value:  0.000124014711404701"
```

Next, the report uses backward selection to select some statistically significant variables. Results of **Model 2**:

```
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
## [1] "Residual Standard Error:  2.45884648868451"
## [1] "Degrees of Freedom:  28"
## [1] "Adjusted R-squared:  0.833556080257604  (model can explain 83% of the variance of the MPG variable)"
## [1] "p-value:  1.21044643016211e-11"
```

Next, the report fits the simple model with MPG as the outcome variable and Transmission as the predictor variable. It shows that on average, a car has 17.147 mpg with automatic transmission, and if it is manual transmission, 7.245 mpg is increased. Results of **Model 3**:

```
## lm(formula = mpg ~ am, data = mtcars)
## [1] "Residual Standard Error:  4.90202882893156"
## [1] "Degrees of Freedom:  30"
## [1] "Adjusted R-squared:  0.338458908206314  (model can explain 34% of the variance of the MPG variable)"
## [1] "p-value:  0.000285020743935068"
```

Lastly, according to the scatter plot (*Figure 3*), it indicates that there appears to be an interaction between "wt" and "am" variables, since automatic cars tend to weigh heavier than manual cars. Results of **Model 4**:

```
##
## 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 ***
## am1           14.079      3.435   4.099 0.000341 ***
## wt:am1        -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
```

The report selects the final model (*Model 4*) with the highest Adjusted R-squared value (88%), "mpg ~ wt + qsec + am + wt:am". Model 4 results show that when "wt" (weight lb/1000) and "qsec" (1/4 mile time) remain constant, cars with manual transmission add 14.079 + (-4.141)*wt more MPG (miles per gallon) on average than cars with automatic transmission. That is, a manual transmitted car that weighs 2000 lbs have 5.797 more MPG than an automatic transmitted car that has both the same weight and 1/4 mile time.

Residual Analysis and Diagnostics According to the residual plots (*Figure 4*), we can verify the following underlying assumptions:
1. The Residuals vs. Fitted plot shows no consistent pattern, supporting the accuracy of the independence assumption.
2. The Normal Q-Q plot indicates that the residuals are normally distributed because the points lie closely to the line.
3. The Scale-Location plot confirms the constant variance assumption, as the points are randomly distributed.
4. The Residuals vs. Leverage argues that no outliers are present, as all values fall well within the 0.5 bands.

As for the Standardized Difference of the Beta, the measure of how much an observation has effected the estimate of a regression coefficient. Results below prove that the report meets all basic assumptions of linear regression model.

```
## [1] "dfbetas:  0"
```

Figure 1 - Boxplot of MPG vs. Transmission

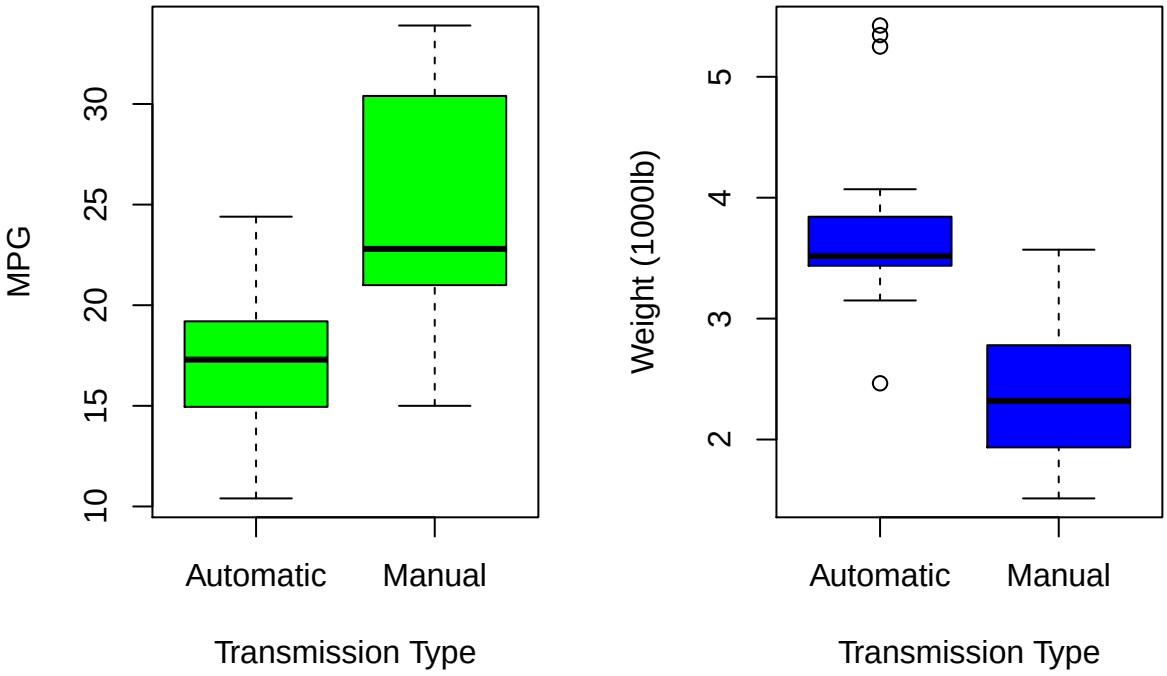


Figure 2 - Pair Graph of Motor Trend Dataset

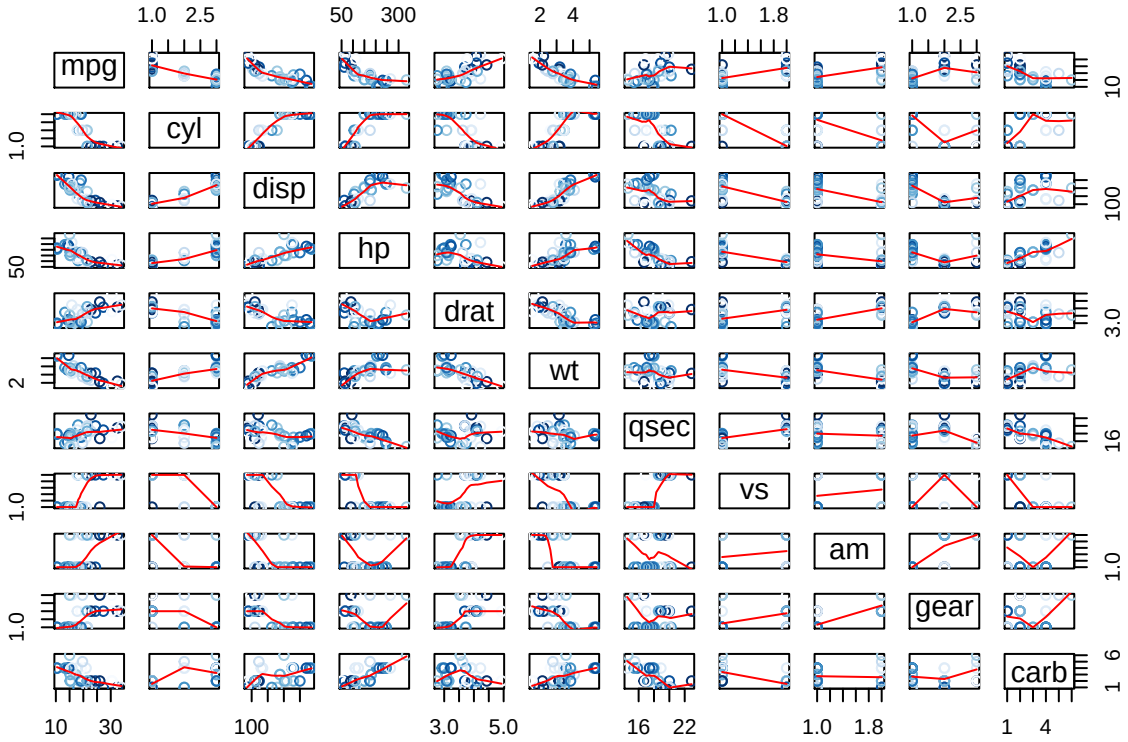


Figure 3 - Scatter Plot of MPG vs. Weight by Transmission

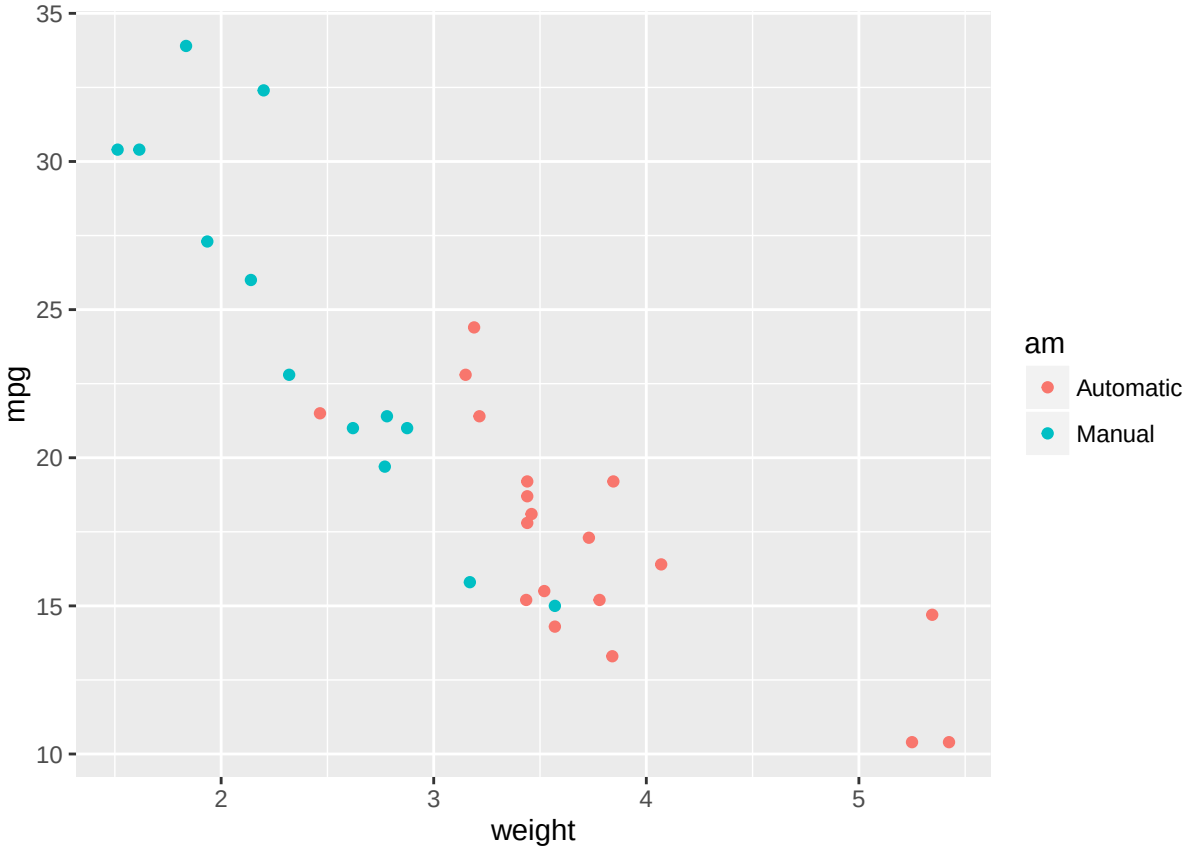


Figure 3(a) - Scatter Plot of MPG vs. Cylinder

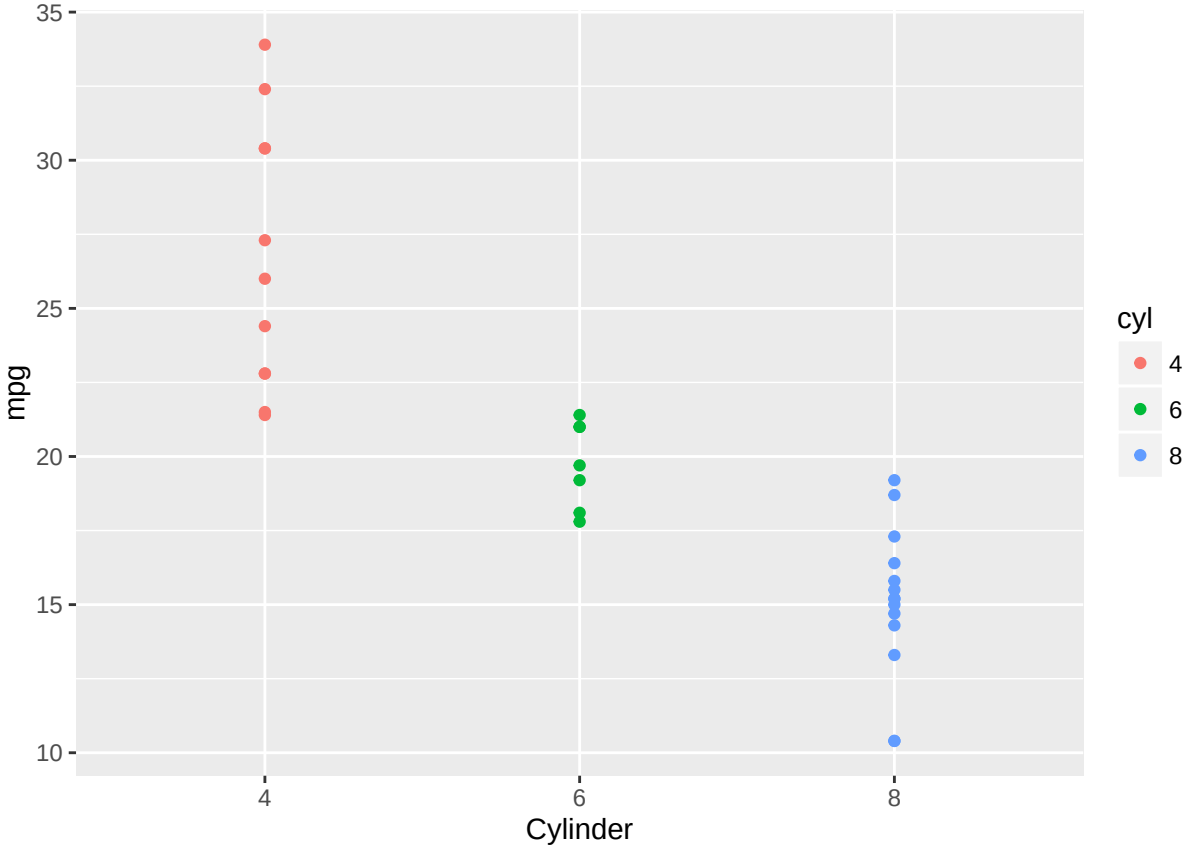
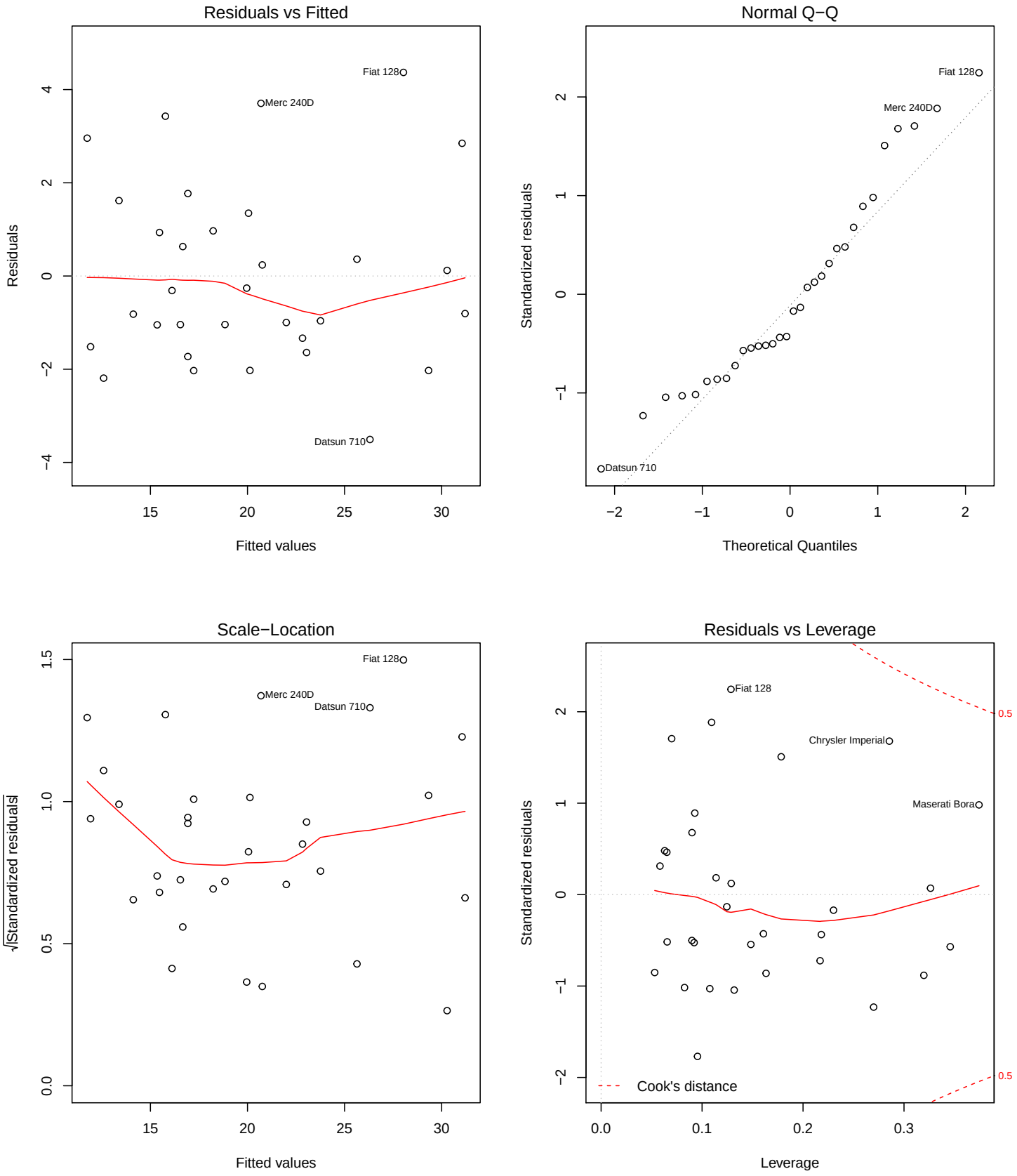


Figure 4 - Residual Plots



GitHub Repo