

# mtcars analysis (extension)

*Pascal P*

*24 August 2018*

## Contents

|          |   |          |
|----------|---|----------|
| <b>1</b> | <b>Summary</b>                                    | <b>1</b> |
| <b>2</b> | <b>Analysis (alternative model)</b>               | <b>2</b> |
| 2.1      | Preparation . . . . .                             | 2        |
| 2.2      | Model . . . . .                                   | 2        |
| 2.3      | Regression diagnostics . . . . .                  | 2        |
| 2.3.1    | Multiplots . . . . .                              | 2        |
| 2.3.2    | Going further with Normality assumption . . . . . | 3        |
| 2.3.3    | Transforming the response variable . . . . .      | 4        |
| 2.3.4    | Unusual observation . . . . .                     | 4        |
| 2.4      | Comparisons . . . . .                             | 5        |
| <b>3</b> | <b>Conclusion</b>                                 | <b>5</b> |

## 1 Summary

This document is an (optional investigative) extension of another study `analysis_mtcars.Rmd` (located in the same directory).

Having noticed a possible interaction between `wt` and `am` in the `mtcars` dataset, I would like to present this alternative model and compare it with initial model (in `analysis_mtcars.Rmd` document).

In the following, I am going to present the alternative model, investigate the regression diagnostics and then present a comparison before presenting the conclusions which state what I believe was achieved.

## 2 Analysis (alternative model)

### 2.1 Preparation

The dataset (`mtcars`) once prepared, meaning all numerical (discrete) values (for features `cyl`, `vs`, `gear`, `carb` and `am`) converted as factor look as followed (extract):

|                   | 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  | manual | 4    | 4    |
| Mazda RX4 Wag     | 21.0 | 6   | 160  | 110 | 3.90 | 2.875 | 17.02 | 0  | manual | 4    | 4    |
| Datsun 710        | 22.8 | 4   | 108  | 93  | 3.85 | 2.320 | 18.61 | 1  | manual | 4    | 1    |
| Hornet 4 Drive    | 21.4 | 6   | 258  | 110 | 3.08 | 3.215 | 19.44 | 1  | auto   | 3    | 1    |
| Hornet Sportabout | 18.7 | 8   | 360  | 175 | 3.15 | 3.440 | 17.02 | 0  | auto   | 3    | 2    |
| Valiant           | 18.1 | 6   | 225  | 105 | 2.76 | 3.460 | 20.22 | 1  | auto   | 3    | 1    |

### 2.2 Model

Let us consider, the interaction between `wt` and `am` (as noted on our first plots for exploratory data analysis in document `analysis_mtcars.Rmd`).

```
library(car)
best.model_with_interaction <- lm(mpg ~ wt + qsec + am + wt:am, data = mtcars)

summary(best.model_with_interaction)$coefficient
```

|             | 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 |
| ammanual    | 14.079428 | 3.4352512  | 4.098515  | 0.0003408693 |
| wt:ammanual | -4.141376 | 1.1968119  | -3.460340 | 0.0018085763 |

```
summary(best.model_with_interaction)$r.squared
```

```
[1] 0.8958514
```

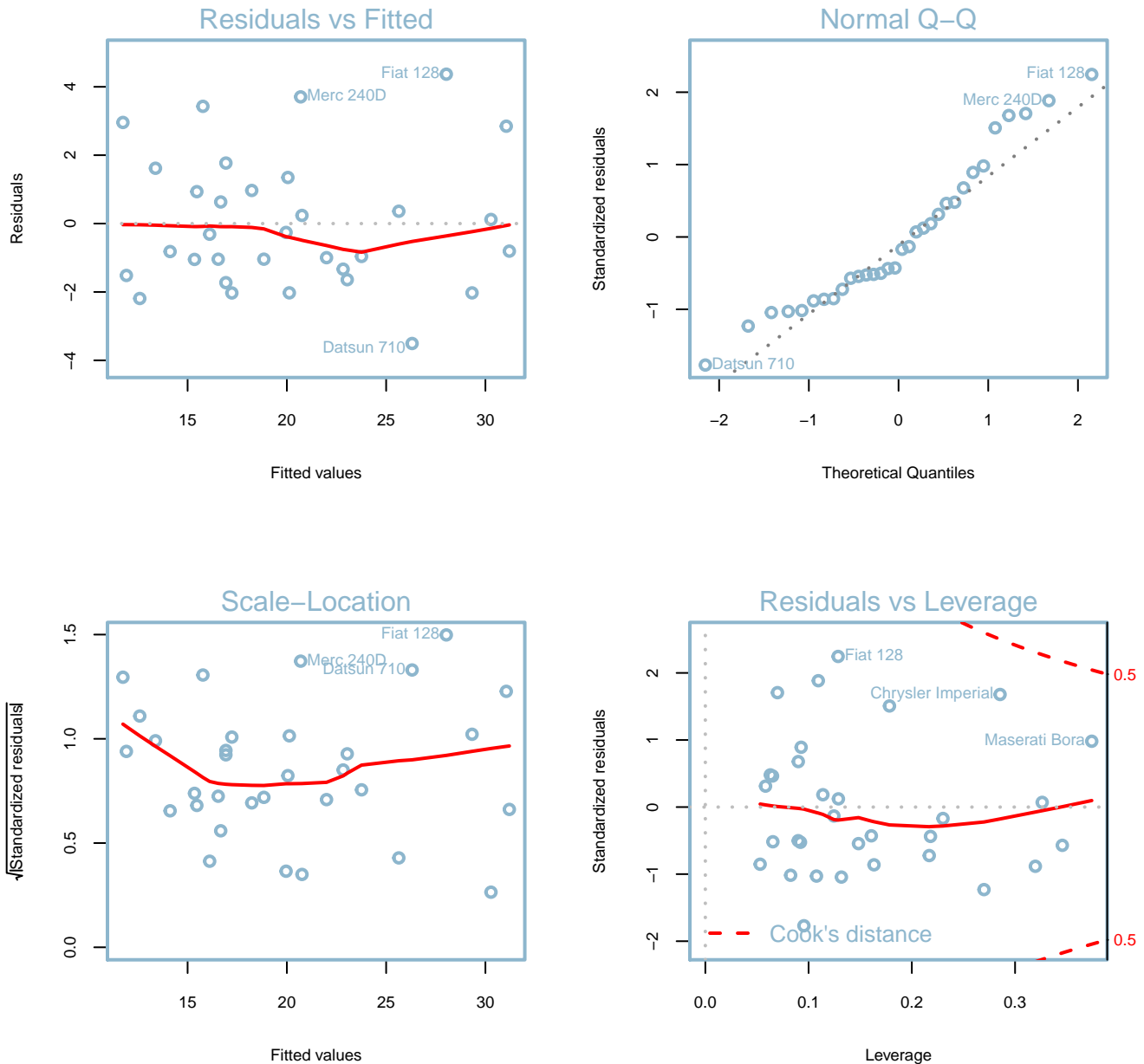
All estimates in this model have small p-value and are therefore significantly different from zero. The percentage of variance explained (linked to  $R^2$  statistic) is about 89.6%.

When holding `wt` and `qsec` constant, manual transmission cars have on average  $14.079 - 4.141 = 9.938$  more `mpg` than automatic transmission ones.

### 2.3 Regression diagnostics

#### 2.3.1 Multiplots

The standard plots for residual diagnostics are presented below:



We observe that:

- there is no particular pattern for **Residuals vs fitted** (assumption of linearity), **Scale-Location** (assumption of homoscedasticity) plots,
- we may have in **Residuals vs Leverage** plots (unusual observations), some points to look at (for later).
- Independence of the dependent variable values (**mpg**) can be assumed given the dataset (my assumption, no reason to believe that the characteristics of one car model affects directly another one).
- however it is clear that the assumption of normality (on the residuals) does not hold (**Normal Q-Q** plot), this may suggest:
  - 1 - that we need to further assess the robustness of that assumption (with **Shapiro-Wilk normality test**) or alternatively
  - 2 - we need to transform the response variable **mpg** (using **powerTransform** function from **car** package).

### 2.3.2 Going further with Normality assumption

Let's try the **Shapiro-Wilk normality test** on the residuals:

```
shapiro.test(best.model_with_interaction$res)
```

Shapiro-Wilk normality test

```
data: best.model_with_interaction$res  
W = 0.94444, p-value = 0.1001
```

And here we would conclude that the normality assumption (on the residuals) is holding (p-value of 0.1).

### 2.3.3 Transforming the response variable

The alternative suggested was to transform the response variable (denoted  $Y$ , to  $Y^\lambda$ . Let's check this with the following code:

```
summary(powerTransform(mtcars$mpg))
```

bcPower Transformation to Normality

|             | Est    | Power | Rounded | Pwr | Wald    | Lwr | Bnd | Wald   | Upr | Bnd |
|-------------|--------|-------|---------|-----|---------|-----|-----|--------|-----|-----|
| mtcars\$mpg | 0.0296 |       | 1       |     | -1.0107 |     |     | 1.0698 |     |     |

Likelihood ratio test that transformation parameter is equal to 0  
(log transformation)

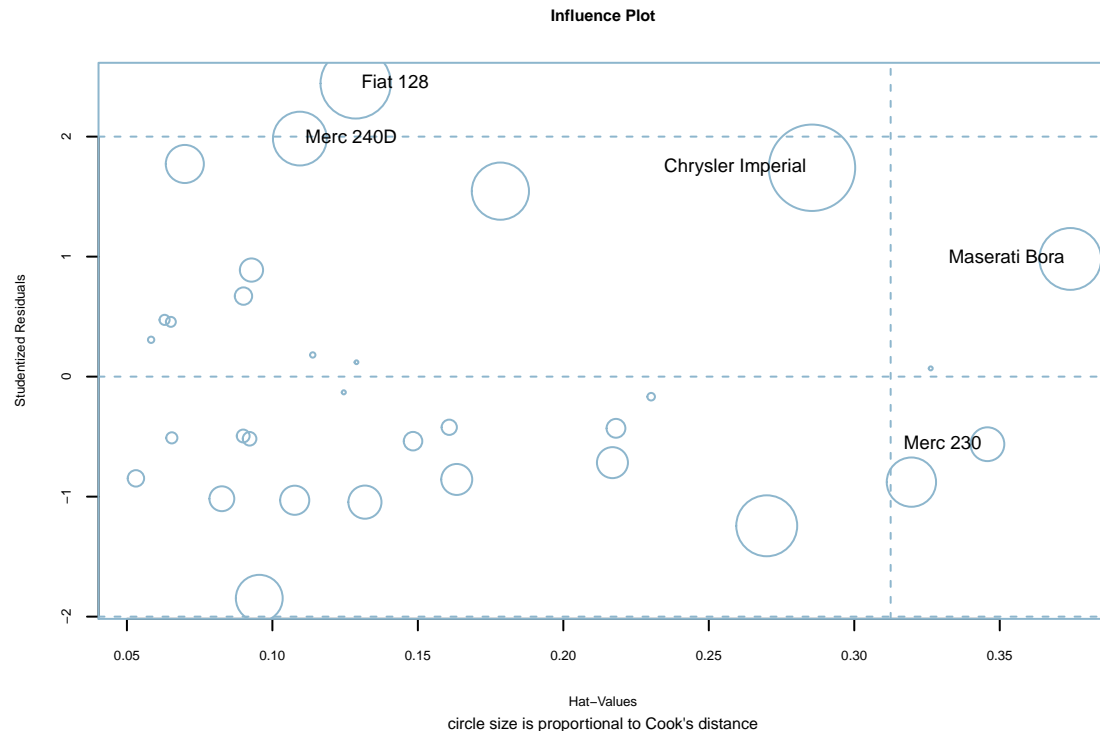
|                       | LRT        | df | pval    |
|-----------------------|------------|----|---------|
| LR test, lambda = (0) | 0.00310595 | 1  | 0.95556 |

Likelihood ratio test that no transformation is needed

|                       | LRT      | df | pval    |
|-----------------------|----------|----|---------|
| LR test, lambda = (1) | 3.212664 | 1  | 0.07307 |

A log transformation is suggested although the non-transformation hypothesis ( $\lambda = 1$ ) cannot be rejected (p-value of 0.07), which means that there is not so much evidence that a transformation is required.

### 2.3.4 Unusual observation



|                   | StudRes    | Hat       | CookD     |
|-------------------|------------|-----------|-----------|
| Merc 240D         | 1.9835159  | 0.1094330 | 0.0872122 |
| Merc 230          | -0.5632192 | 0.3457671 | 0.0344001 |
| Chrysler Imperial | 1.7404165  | 0.2854605 | 0.2251060 |
| Fiat 128          | 2.4433618  | 0.1286155 | 0.1488367 |
| Maserati Bora     | 0.9806330  | 0.3742640 | 0.1151987 |

We observe:

- `Fiat 128` is an outlier,
- `Merc 230` and `Maserati Bora` have high leverage and
- and observations with relative large circle may have disproportionate influence.

## 2.4 Comparisons

Let's compare this model with the one presented in original analysis (document `analysis_mtcars.Rmd`), with `anova` and AIC (Akaike Information Criterion) functions. First let's show the previous model, as followed:

```

      Estimate Std. Error  t value    Pr(>|t|)
(Intercept)  9.617781   6.9595930   1.381946 1.779152e-01
wt          -3.916504   0.7112016  -5.506882 6.952711e-06
qsec         1.225886   0.2886696   4.246676 2.161737e-04
ammanual     2.935837   1.4109045   2.080819 4.671551e-02
[1] 0.8496636

```

The following comparison tells us that our model with interaction is indeed better (small p-value) by adding the interaction term.

```
anova(prev_best.model, best.model_with_interaction)
```

| Res.Df | RSS      | Df | Sum of Sq | F        | Pr(>F)    |
|--------|----------|----|-----------|----------|-----------|
| 28     | 169.2859 | NA | NA        | NA       | NA        |
| 27     | 117.2763 | 1  | 52.00966  | 11.97396 | 0.0018086 |

With the following comparison we look for the model with the smallest AIC values, and once again it is suggested that adding the interaction term is relevant.

```
AIC(prev_best.model, best.model_with_interaction)
```

|                             | df | AIC      |
|-----------------------------|----|----------|
| prev_best.model             | 5  | 154.1194 |
| best.model_with_interaction | 6  | 144.3736 |

## 3 Conclusion

- Interpreting the results correctly looks more like an art (with science) than a science
- The size of the sample (32 observations) is quite small and even more in comparison with the number of features (11).  
This small sample size may limit the power of our analysis and the accuracy of our conclusion
- The alternative model with interaction term seems better (than original model) and almost compatible with all the related assumptions of a multi-linear regression model: Independence, Linearity, Homoscedasticity (constant variance), and Normality (although not clear cut to me for this one).

- If we accept the previous point, we can state that:

1. the alternative model does not change the main claim of the original one, namely that manual transmission cars have a better `mpg` (lower consumption) than automatic ones (a 2.94 more mpg vs 9.94 more mpg with the alternative model).
2. however the alternative model (with interaction term between `wt` (weight and `am` transmission)) is slightly better.

Would this hold on a larger dataset? would we be able to better select alternative models and use cross-validation to quantify how better models are? It looks like we need, as usual more data (if not more features).