Analysis of the effect of the transmission on the MPG

Erika R. Frits

2018-03-11

Summary

The aim of the analysis is to answer the following questions:
- Is an automatic or manual transmission better for MPG
- Quantify the MPG difference between automatic and manual transmissions

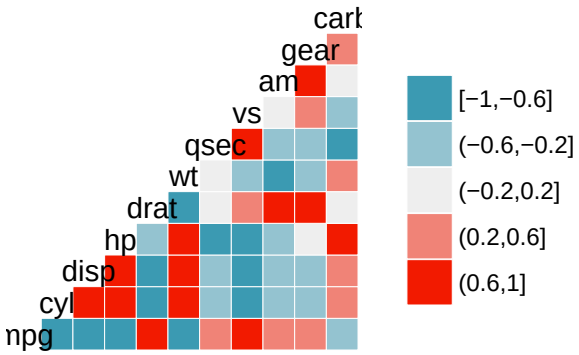
I selected the model $mpg_i = b_0 + b_1 * transmission_i + b_2 * hp_i + e_i$.

The anaylsis showed that the cars with manual transmission can run 4.0681 $\pm$ 1.2587 miles per gallon compared to cars with automatic transmission, if the compared cars have the same HP.

Exploring the data

After checking the dataset I found that
- all variables are coded as numeric values
- there are no missing values in the dataset Note: V/S is the engine type (0 = V-engine, 1 = stright engine)

Correlations between the variables:



Transmission has moderate correlation with mpg. The variables with the highest correlation to mpg are the number of cylinders, the weight, the displacements and horse power. These variables are also highly correlated with each other. the horse power has the lowest correlation with transmission. The other variables has moderate correlation with mpg, but the number of forward greas are highly correlated to transmission and the number of carburators and the 1/4 mile time to horse power.

Detailed plots are in the Appendix 1.

I converted the follwing variables to factors:
- vs (engine): {0 = 'V', 1 = 'S'}
- am (transmission): {0 = 'A', 1 = 'M'}

I could also convert 'cyl', 'gear', 'carb' to factors, but I left them as numbers for the simplicity of the model.

Model selection

The simpliest model: mpg vs. transmission

I start with the simpliest model, which contains only the *transmission* variable: $mpg_i = b_0 + b_1 * transmission_i + e_i$

```
              Estimate Std. Error   t value     Pr(>|t|)
(Intercept)   17.147368   1.124603 15.247492 1.133983e-15
transmissionM  7.244939   1.764422  4.106127 2.850207e-04
```

$R^2 = 0.3597989$

The *(Intercept)* coefficient means the MPG value of a theoretical car with zero HP and automatic transmission. The *transmissionM* coefficient means the change of MPG if the transmission id changed from automatic to manual.

Residual plots are in Appendix 2.

The residuals are evenly spread around 0 which means no systematic error in the model, and according to the Q-Q plot they are normally distributed. The variance of the data is higher in the 'M' transmission group. There are also some possible outlier points which standard resiuals are very close or a little bit over the boundaries of 95% confidence interval (approx. $\pm$ 1.91).

The outlier points: 20, 31

Because of their low leverage and closeness to the other values, they have almost no effect on the coefficients.

The low R^2 value indicates that some other factors has to be taken into account.

Introducing other variables into the model

Note: becuase the outlier points can vary from model to model, I always use the original data set. I will include horse power (*hp*) into the model because according to the correlation matrix the other variables correlate with it or trasnmission.

Adding the horse power

$mpg_i = b_0 + b_1 * transmission_i + b_2 * hp + e_i$

```
               Estimate  Std. Error   t value     Pr(>|t|)
(Intercept)   26.5849137 1.425094292 18.654845 1.073954e-17
transmissionM  5.2770853 1.079540576  4.888270 3.460318e-05
hp            -0.0588878 0.007856745 -7.495191 2.920375e-08
```

$R^2 = 0.7820346$

The R^2 value became higher, than the model without *hp* and the t-test also find it significant.

Residual plots can be found in Appendix 3

The residuals does not indicate any pattern, but there are some possible outliers and one point with very high leverage. The Q-Q plot also shows some points at the lower end which do not fit on the line.

The found outlier points: 20, 29, 31

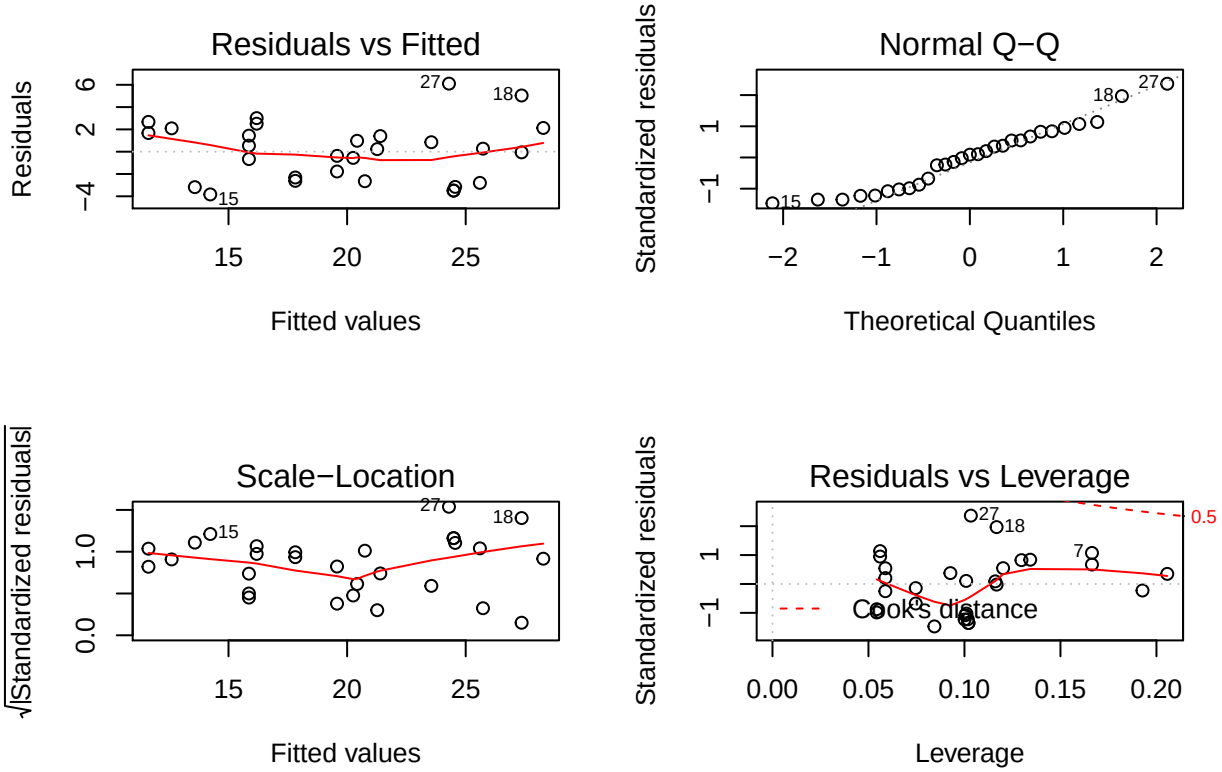
The coefficients leaving out the outlier points:

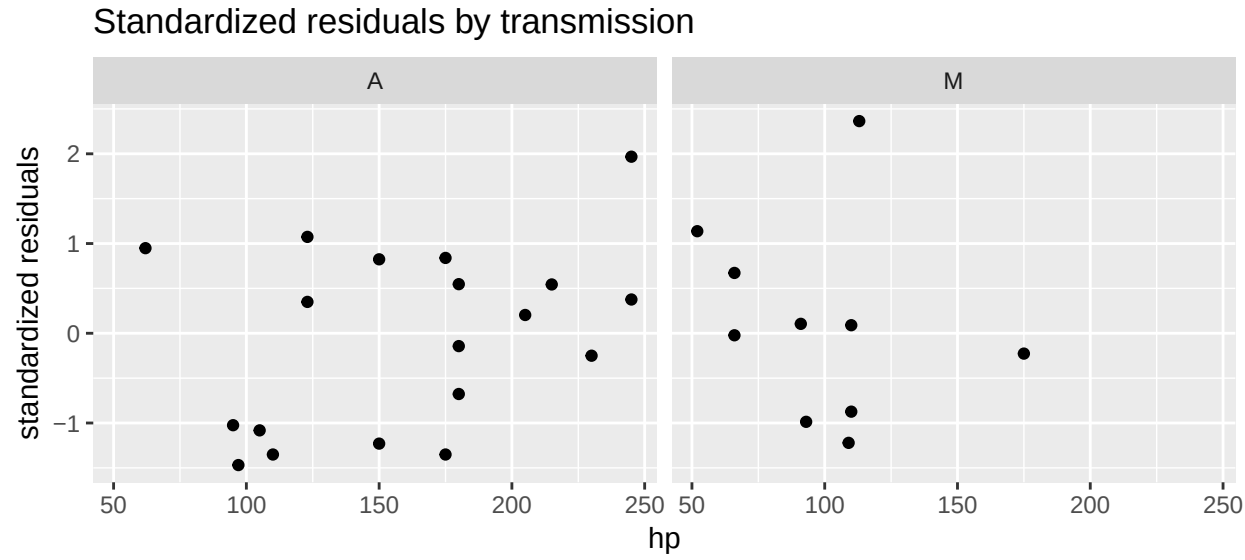
```
                Estimate Std. Error   t value     Pr(>|t|)
(Intercept)   27.59013499 1.84865369 14.924448 2.900273e-14
transmissionM  4.06813681 1.25871077  3.231987 3.327730e-03
```

```
hp          -0.06516012 0.01085459 -6.003002 2.442996e-06
```

$R^2 = 0.7828669$

Checking the residuals:





Leaving the outliers out made the Q-Q plot's lower tail more reamarkable. I tested the normality of the standard residuals with Shapiro-Wilks test.

The p-value (0.2138414) is higher than 0.05 so these points are coming from a normally distributed dataset.

Conclusion

I selected the model $mpg_i = b_0 + b_1 * transmission_i + b_2 * hp_i + e_i$.

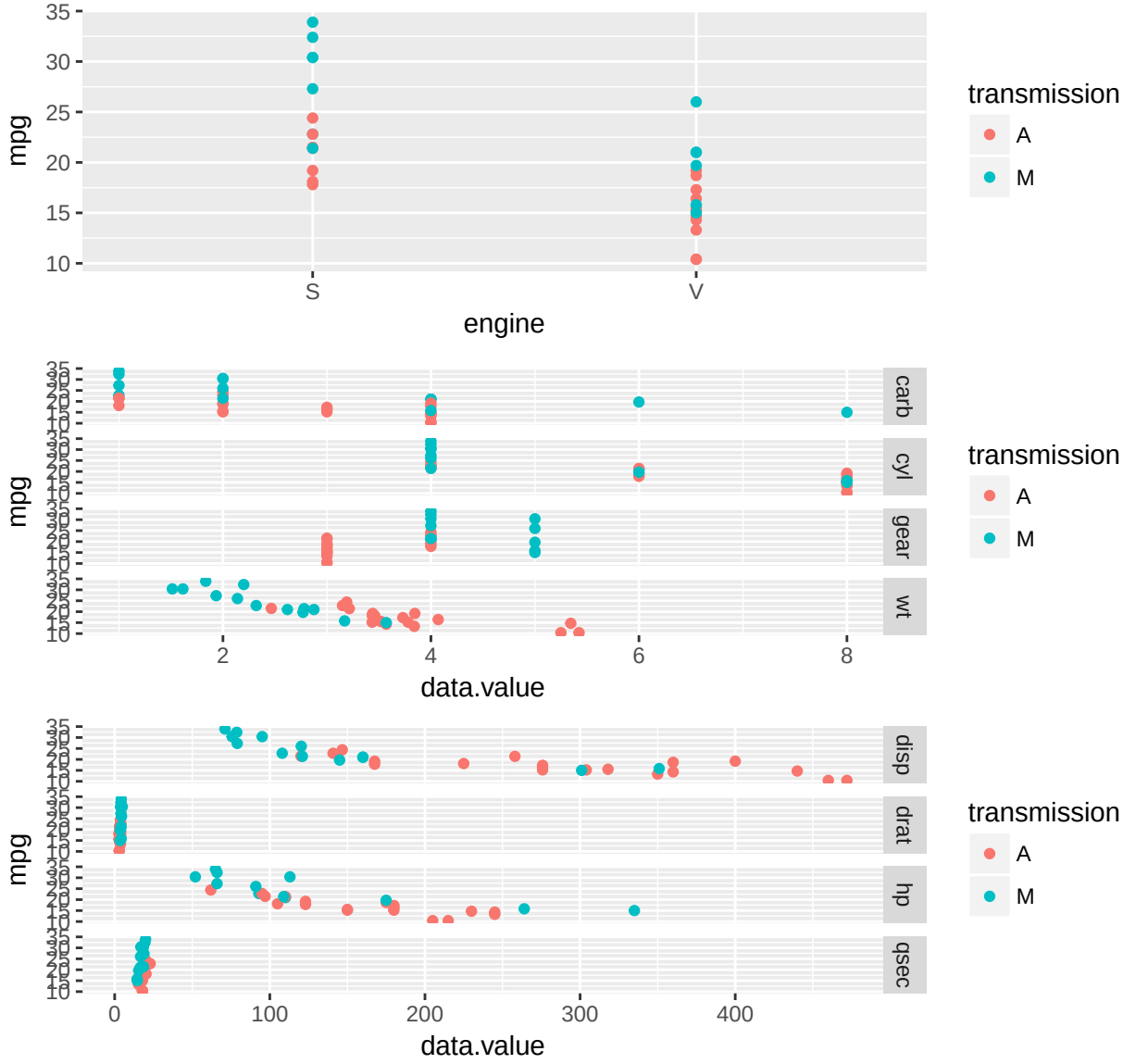
The final coefficients (leaving the outliers out):

```
Call:
lm(formula = mpg ~ transmission + hp, data = carData.m1)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8323 -2.6161  0.2304  1.6741  6.1048

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   27.59013    1.84865  14.924 2.90e-14 ***
transmissionM  4.06814    1.25871   3.232  0.00333 **
hp            -0.06516    0.01085  -6.003 2.44e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.727 on 26 degrees of freedom
Multiple R-squared:  0.7829,    Adjusted R-squared:  0.7662
F-statistic: 46.87 on 2 and 26 DF,  p-value: 2.385e-09
```

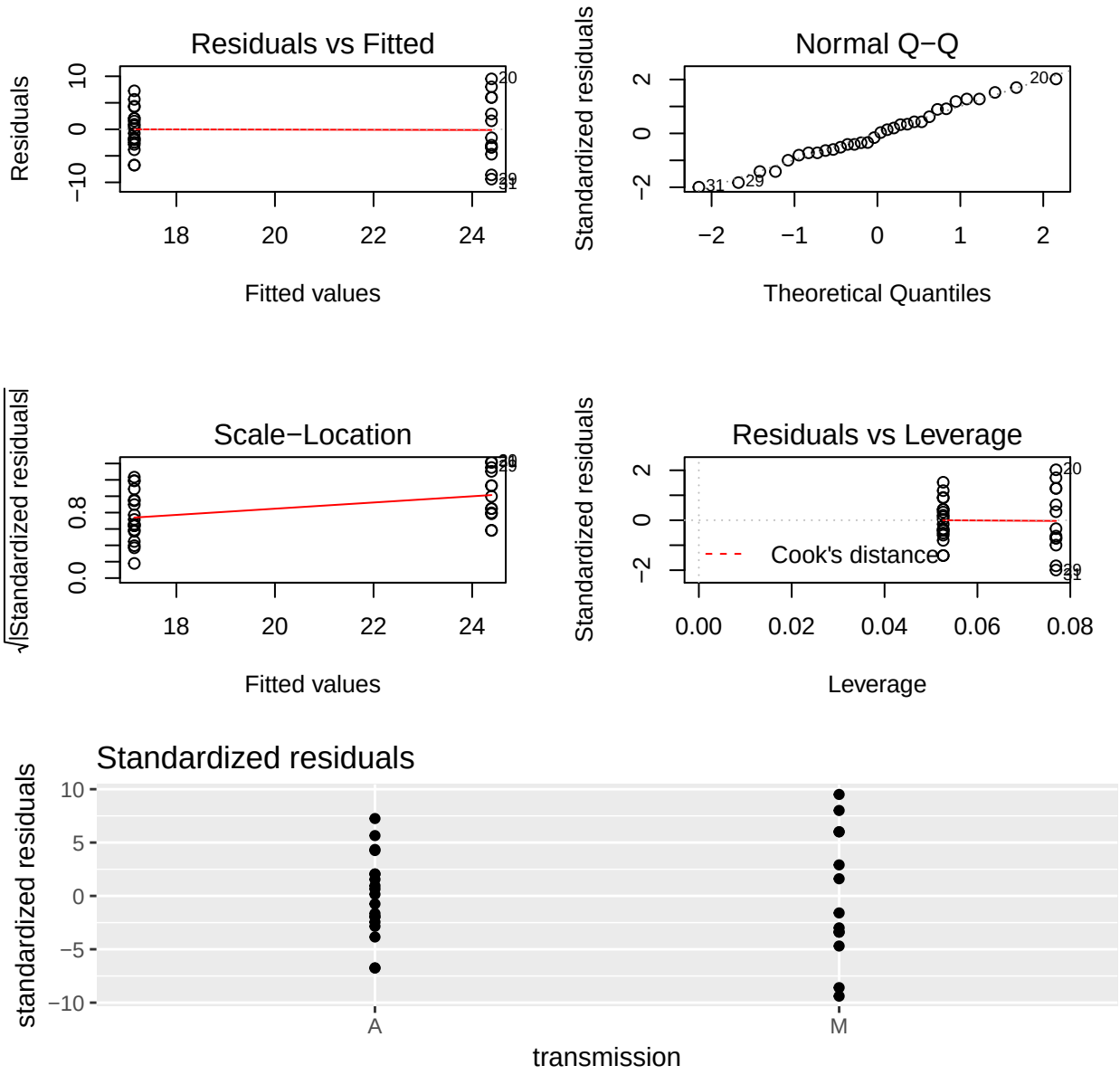
This anaylsis showed that the cars with manual transmission can run 4.0681 $\pm$ 1.2587 miles per gallon compared to cars with automatic transmission, if the compared cars have the same HP.

Fun fact: if this model were right for any HP, then a car with manual transmission and 485.9 HP could not move... :)

Appendix 1 : the data



Appendix 2: residuals of the simple model



Appendix 3: residuals of the model with horse power term