Regression Models - Project

Loïc BERTHOU

August 21, 2015

Executive Summary

In this document, we will explore the question “Is an automatic or manual transmission better for MPG” by analysing the data provided in the **mtcars** dataset from R. To do so we have performed some basic exploratory analysis that will help us direct our more detailed data analysis. It will be demonstrated that the manual transmission is better for MPG in general but that other factors have a great influence on MPG.

Exploratory Analysis

In the exploration of this dataset, we first notice that all columns are coded as numbers, despite the fact that some parameters clearly represent factors (**am** and **vs**). We could argue that some other parameters could be considered as factors (**cyl**, **gear**, **carb**). We will keep this in mind while performing our analysis and force some conversions if needed.

This dataset being quite small, it is possible to view it entirely and we won’t need the *summary* or *str* function to have an overview. It could be interesting to plot some relationships right away but to do an unbiased analysis, we will first look at the correlation matrix (Appendix 1).

It appears clearly that some of the parameters are strongly correlated. We will then try to find the most influential parameters, leaving the correlated ones out of the model.

For a first idea of the influence of weight on the mileage, we will draw a first plot that takes into account the type of transmission (Appendix 2).

Model Selection

To find a model that fits our data, we will first perform the linear regression taking into account all the parameters provided in our dataset. By looking at the coefficients, we should confirm the first intuitions on our strongest influential parameters (Appendix 3).

We already know that weight (**wt**) is one of the most influential parameter on the mileage. Since the displacement (**disp**) and the number of cylinders (**cyl**) and the gross horsepower (**hp**) are strongly correlated to weight, I will leave them out of the model.

By removing the weakest parameters successively, and with the analysis of the variance between the various models (Appendix 4) we ended up with the model *fitSimpleQsec* that was reasonably simple and yet quite accurate.

The residual plots indicate a good fit (Appendix 5).

Conclusion

We can certainly say that the choice of the transmission has an impact on the mileage. Our model shows that for all other parameters being equal, the manual transmission has 14 MPG more than the automatic.

However, we also showed that the weight of the car and its performance can have a significant impact on the mileage.

It is also important to note that the number of cars used for this analysis is yet limited and this should be investigated further with a more important number of cars for more significant results. We could also add other parameters that might influence the mileage (gasoline/diesel, car make, sedan/wagon, etc).

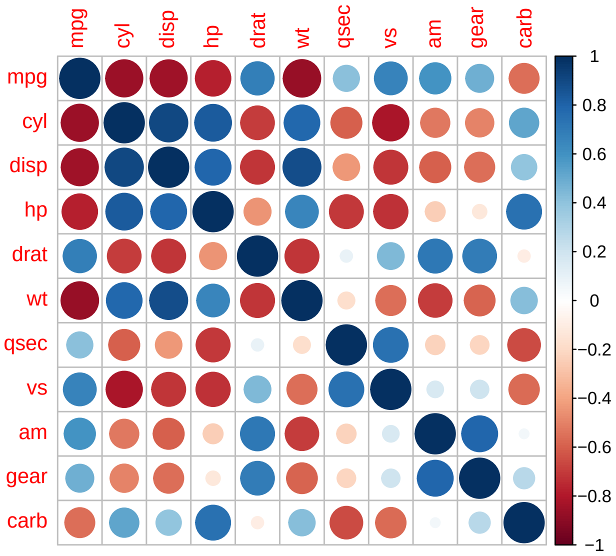
Appendix

Appendix 1. Correlation matrix

```
library(corrplot)
corMtcars <- cor(mtcars)
round(corMtcars, digits=2)
```

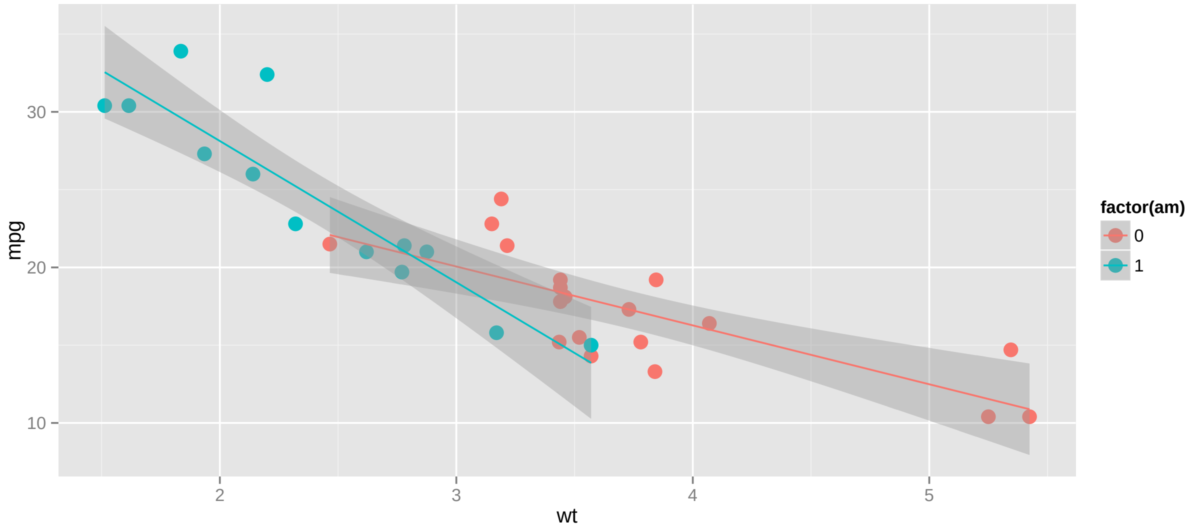
```
##        mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
## mpg   1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
## cyl  -0.85  1.00  0.90  0.83 -0.70  0.78 -0.59 -0.81 -0.52 -0.49  0.53
## disp -0.85  0.90  1.00  0.79 -0.71  0.89 -0.43 -0.71 -0.59 -0.56  0.39
## hp   -0.78  0.83  0.79  1.00 -0.45  0.66 -0.71 -0.72 -0.24 -0.13  0.75
## drat  0.68 -0.70 -0.71 -0.45  1.00 -0.71  0.09  0.44  0.71  0.70 -0.09
## wt   -0.87  0.78  0.89  0.66 -0.71  1.00 -0.17 -0.55 -0.69 -0.58  0.43
## qsec  0.42 -0.59 -0.43 -0.71  0.09 -0.17  1.00  0.74 -0.23 -0.21 -0.66
## vs    0.66 -0.81 -0.71 -0.72  0.44 -0.55  0.74  1.00  0.17  0.21 -0.57
## am    0.60 -0.52 -0.59 -0.24  0.71 -0.69 -0.23  0.17  1.00  0.79  0.06
## gear  0.48 -0.49 -0.56 -0.13  0.70 -0.58 -0.21  0.21  0.79  1.00  0.27
## carb -0.55  0.53  0.39  0.75 -0.09  0.43 -0.66 -0.57  0.06  0.27  1.00
```

```
corrplot(corMtcars)
```



Appendix 2. Relationship between wt, wt and am

```
g <- ggplot(mtcars, aes(x=wt, y=mpg, color=factor(am)))
g <- g + geom_point(size=4)
g <- g + geom_smooth(method="lm")
g
```



Appendix 3. Linear Regression including all parameters

```
fitAll <- lm(mpg ~ ., data=mtcars)
summary(fitAll)$coef
```

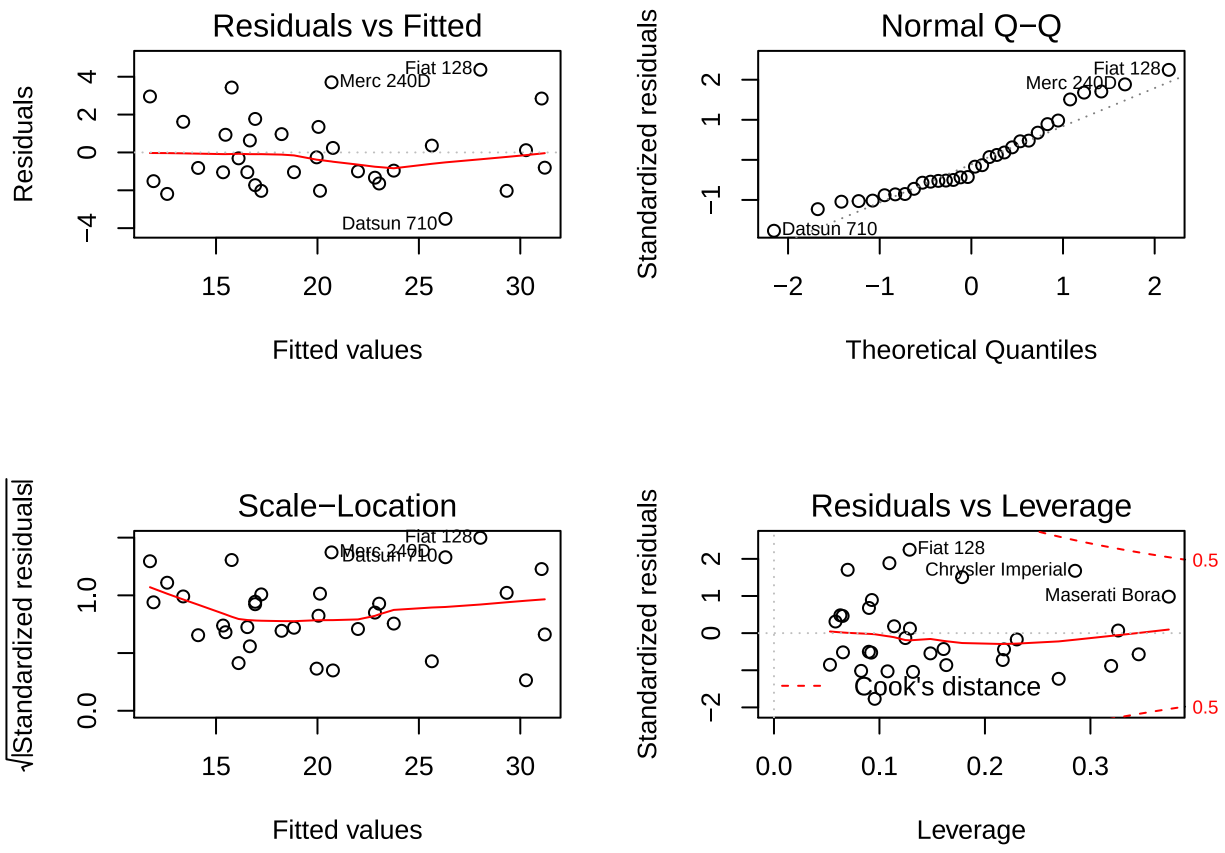
```
##                Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept) 12.30337416 18.71788443  0.6573058 0.51812440
## cyl         -0.11144048  1.04502336 -0.1066392 0.91608738
## disp         0.01333524  0.01785750  0.7467585 0.46348865
## hp          -0.02148212  0.02176858 -0.9868407 0.33495531
## drat         0.78711097  1.63537307  0.4813036 0.63527790
## wt          -3.71530393  1.89441430 -1.9611887 0.06325215
## qsec         0.82104075  0.73084480  1.1234133 0.27394127
## vs           0.31776281  2.10450861  0.1509915 0.88142347
## am           2.52022689  2.05665055  1.2254035 0.23398971
## gear         0.65541302  1.49325996  0.4389142 0.66520643
## carb        -0.19941925  0.82875250 -0.2406258 0.81217871
```

Appendix 4. Model selection with analysis of variance

```
fitSimple <- lm(mpg ~ wt * factor(am), data=mtcars)
fitSimpleQsec <- lm(mpg ~ wt * factor(am) + qsec, data=mtcars)
fitComplex <- lm(mpg ~ wt * factor(am) + qsec + factor(vs) + gear + carb + drat, data=mtcars)
anova(fitSimple, fitSimpleQsec, fitComplex)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt * factor(am)
## Model 2: mpg ~ wt * factor(am) + qsec
## Model 3: mpg ~ wt * factor(am) + qsec + factor(vs) + gear + carb + drat
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)    
## 1     28 188.01                                   
## 2     27 117.28  1    70.731 14.4052 0.0009337 ***
## 3     23 112.93  4     4.343  0.2211 0.9238752    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix 5. Model Residuals



Residuals vs Fitted
Residuals
Fitted values
-4
-2
0
2
4
15
20
25
30
Fiat 128
Merc 240D
Datsun 710
Normal Q−Q
Standardized residuals
Theoretical Quantiles
−1
0
1
2
−2
−1
0
1
2
Fiat 128
Merc 240D
Datsun 710
Scale−Location
√|Standardized residuals|
Fitted values
0.0
1.0
15
20
25
30
Fiat 128
Merc 240D
Datsun 710
Residuals vs Leverage
Standardized residuals
Leverage
−2
0
1
2
0.0
0.1
0.2
0.3
Fiat 128
Chrysler Imperial
Maserati Bora
Cook's distance
0.5
0.5