MPG for automatic and manual transmissions

ElisaRMA

6/10/2021

Summary

This report analyses the mtcars dataset with the objective to explore the relationship between miles per gallon (MPG) and other variables, specially the transmission type (manual or automatic). This report was elaborated as part of the Data Science Specialization at Coursera.

Exploratory Data Analysis

The mtcars dataset is a available at R base using the `data()` function as follows. To begin a exploratory data analysis is useful. For such analysis, the function `boxplot` with the two variables of interest can be used, as demonstrated below. For the `am` variable 0 is automatic and 1 is manual and by observing the plot, it is possible to infer that manual cars have a higher `mpg`. (plot in the appendix)

```
data("mtcars")
boxplot(mtcars$mpg ~ mtcars$am)
```

Model Selection

With this in mind, we will analyze the relationship between `mpg` and `am` only, ignoring the other variables, using a linear regression. The formula is, then: mpg = intercept + am*slope

```
model1 <- lm(mpg ~ factor(am), data = mtcars)
summary(model1)$coeff
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## factor(am)1  7.244939   1.764422  4.106127 2.850207e-04
```

As observed, the intercept or mean `mpg` for automatic cars is 17.147. Therefore, we expect an increase of 7.245 in `mpg` for manual cars. However, by looking at the R-squared of 0.3385 (below) it is evident that only 34% of the total variation on `mpg` is represented by the model, so `am` alone is not enough to explain the tendencies in `mpg`. The plot showing this total variation explained by the model is in the appendix.

```
summary(model1)$adj.r.squared
```

```
## [1] 0.3384589
```

With this information, we can attempt to find another model that fits the data in a more efficient way, using the remaining variables of the mtcars dataset and the anova test:

```
model2 <- lm(mpg ~ factor(am)+cyl+disp,data=mtcars)
model3 <- lm(mpg ~ factor(am)+cyl+disp+hp+drat,data=mtcars)
model4 <- lm(mpg ~ factor(am)+cyl+disp+hp+drat+wt+qsec,data=mtcars)
model5 <- lm(mpg ~ factor(am)+cyl+disp+hp+drat+wt+qsec+factor(vs)+gear,data=mtcars)
model6 <- lm(mpg ~ factor(am)+cyl+disp+hp+drat+wt+qsec+factor(vs)+gear+carb,data=mtcars)
anova(model1, model2, model3, model4, model5, model6)[,1:6]
```

```
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 252.08  2    468.82 33.3746 3.015e-07 ***
## 3     26 214.50  2     37.58  2.6756   0.09224 .
## 4     24 149.09  2     65.41  4.6563   0.02120 *
## 5     22 147.90  2      1.19  0.0846   0.91917
## 6     21 147.49  1      0.41  0.0579   0.81218
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By the p-values above, by adding the variables `cyl`, `disp`, `wt`and `qsec` the `mpg` variability can be better explained. The model 4 in this case, would be a good model to explain the `mpg`, as demonstrated by its R squared, with a value of 83% approximately. This r squared value demonstrates that 83% of the variability of `mpg`is explained by the model in question and comparing this plot with the one from model 1 is evident how much the model improved with the addition of these 4 variables.

In addition, the residual plot of such model does not follow any pattern, indicating that such model is indeed applicable to this dataset (plot in appendix).

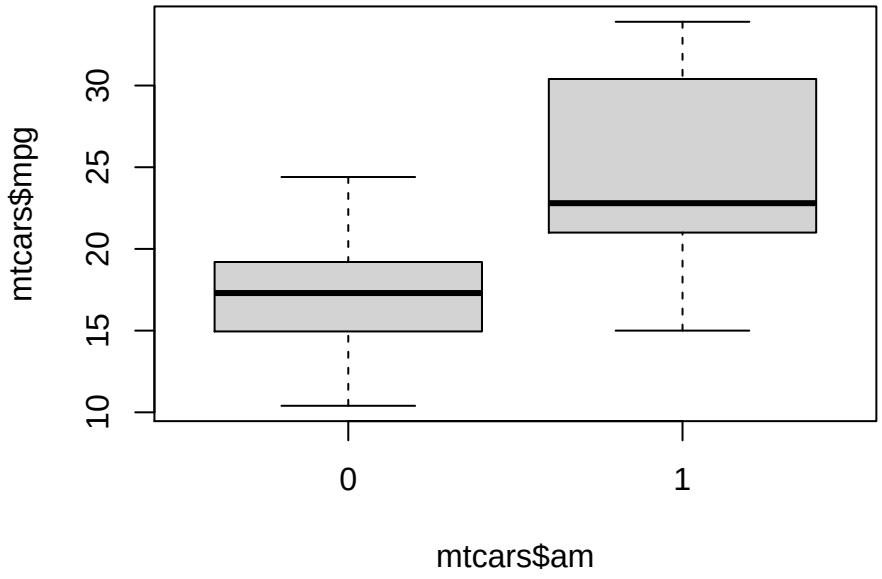
```
summary(model4)$adj.r.squared
```

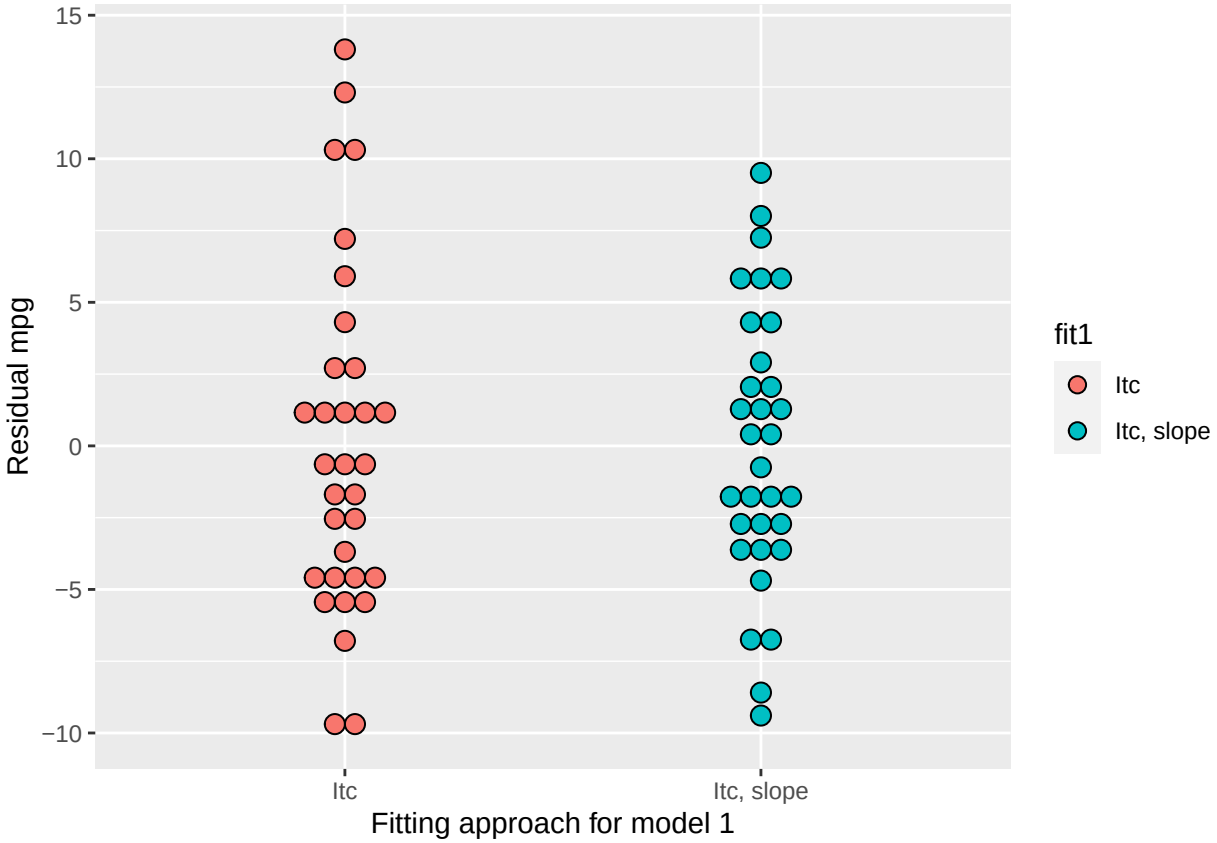
```
## [1] 0.8289819
```

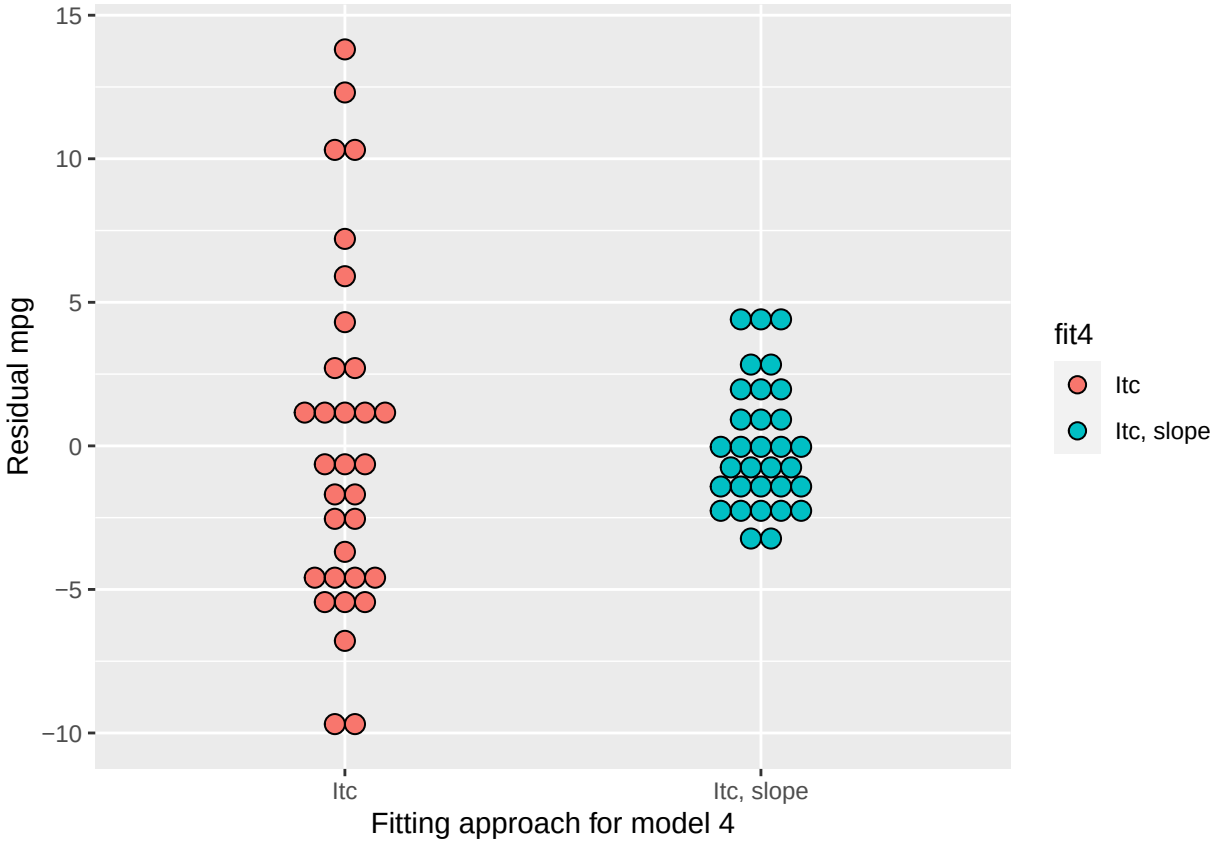
Conclusions

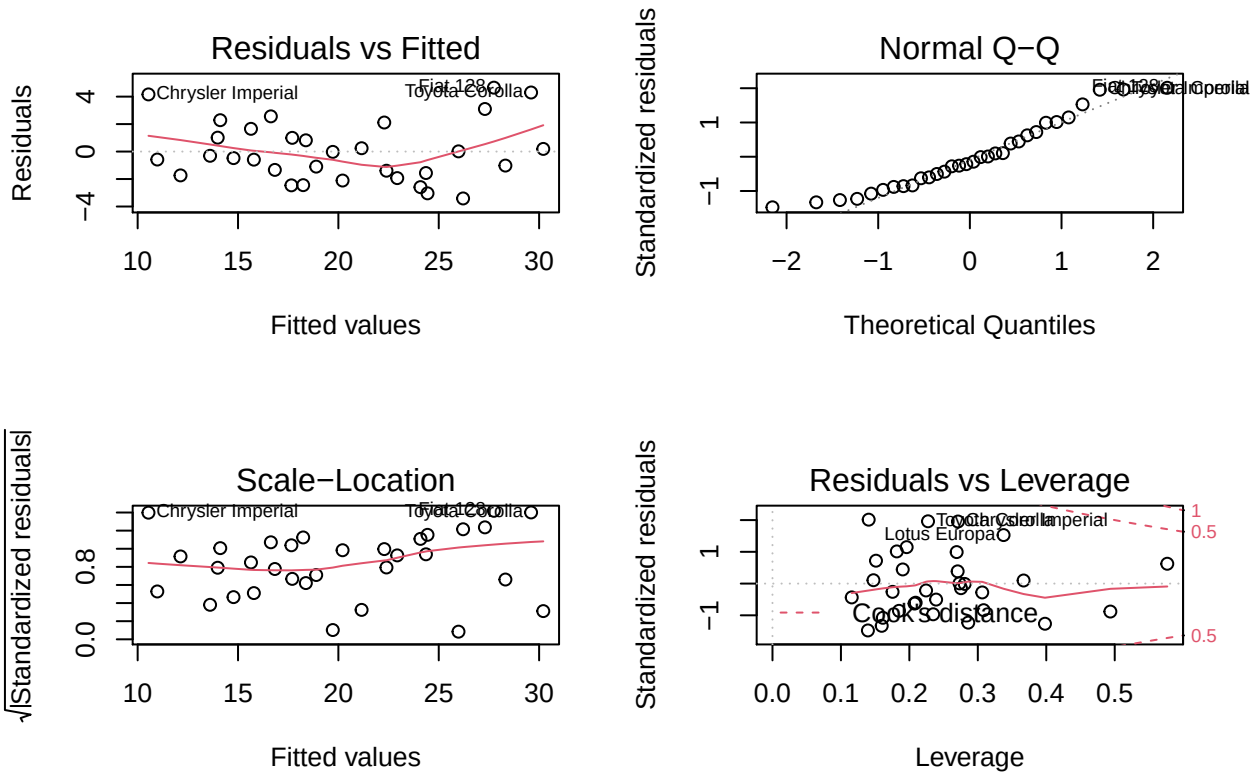
By the analysis developed herein it is possible to conclude that `mpg` can be partially explained by the `am` variable, in which a manual car have a slight tendency to higher `mpg` values (7.25 mpg higher). However, transmission alone is only responsible for 34% of the variability observed in `mpg` and other factors are necessary to explain how many miles/gallon a given car can make.

Appendix


mtcars$mpg
10
15
20
25
30
0
1
mtcars$am


Residual mpg
15
10
5
0
−5
−10
ltc
ltc, slope
Fitting approach for model 1
fit1
ltc
ltc, slope


15
10
5
0
−5
−10
Residual mpg
ltc
ltc, slope
Fitting approach for model 4
fit4
ltc
ltc, slope


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
Lotus Europa