Regression Models Course Project: Auto VS Manual, Which One Is More Fuel Efficient?

Ali Pourkhesalian

07/08/2019

Executive Summary

In this report, first a basic exploratory analysis is performed on the "mtcars" dataset in R, and then, a statistical inferential analysis is carried out. A number of regression models are also checked in order to find the most efficient model. The regression models are compared in terms of complexity and efficiency and it is found out the most important influencing factors on fuel consumption are gearbox type (manual or automatic), the weight of the vehicle and the celerity of the vehicle. Apparently cars with manual transmission consume statistically significantly less fuel than that of automatic cars. However, quantifying the mileage needs considering other influencing variables such as weight and celerity, etc. and adding each variable to the model changes the effect of transmission type on mileage, so quantifying the mileage based on transmission type depends on the number of variable in the model and thus seems to be arbitrary.

Data set

A dataset that is included in R is called "mtcars". As the name implies, the dataset includes data on 32 of cars. The dataset has 11 column which are listed below with a brief description:
1. mpg Miles/(US) gallon, 2. cyl Number of cylinders, 3. disp Displacement (cu.in.), 4. hp Gross horsepower 5. drat Rear axle ratio, 6. wt Weight (1000 lbs), 7. qsec 1/4 mile time, 8. vs Engine (0 = V-shaped, 1 = straight), 9. am Transmission (0 = automatic, 1 = manual), 10. gear Number of forward gears, 11. carb Number of carburetors,

Basic Exploratory Analysis

As mentioned, the dataset has 11 columns/variables and 32 rows/observations. The below table shows a basic exploratory of the data. The table shows the correlation between mpg and other variables in the data. It can be seen that mpg has a strong negative correlation with wt, disp, cyl, hp, and carb, and a positive correlation with am, drat, vs and qsec. It seems that some of the variables that are in strong correlation with mpg are inter-correlated. For example, the displacement volume of the engine of a vehicle is directly proportional to the number of cylinders the vehicle has and so on.

```
##   mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
##  1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
```

Figure 1 in the appendix shows a box plot of mpg related to manual and auto vehicle to see if there is any difference at all. As it can be seen in the figure, type of the gearbox system (automatic or manual) seems to affect the fuel consumption of the vehicle as the box-plot clearly shows that the manual vehicle can travel more miles per every gallon of fuel. Let's check the mean as well. The mean and SEM of mileage for manual cars are 17.15 and 0.68 whereas for automatic cars mean and SEM of mileage are 24.39 and 1.09. However, to make any further inference, one has to carry out a regression analysis on the data to make sure that the difference in the mileage is in fact statistically significant.

Regression Analysis

In this section a few models are fit onto the mtcars dataset and then the models are analysed for efficiency to see how considering the "am" variable in the models can affect the overall efficiency of the model.

Models

Let's first check the simplest model, a linear model of mpg as the response variable and am as the explanatory variable.

```
simp.model <- lm(data= mtcars, mpg~am)
```

The model has an adjusted R square of, 0.34, showing that "am" can explain 34 percent of the variation in "mpg". Now let's find the most efficient model and study it in more details.

```
most.eff.model <- step(lm(mpg ~ ., data = mtcars), direction = "both")
```

Thus, the most parsimonious yet efficient model is the below model:

```
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
```

As it can be seen in the model, "am" is considered an independent explanatory variable. In other word, it definitely has a statistically significant effect on the response variable which in this case is "mpg". The latter model has an R-square of 0.83 meaning that it can explain 83 percent of the variations of mpg. Now, let's compare the two models using anova:

```
anova(simp.model, most.eff.model)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + qsec + am
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     30 720.90
## 2     28 169.29  2    551.61 45.618 1.55e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The above comparison shows that based on the p-value being very small, the latter model is statistically significantly different from the former model and thus the null hypothesis of the two models being the same is rejected. Also, the RSS of the models show that the sum of square residuals in the latter model is less than that of the former.

To quantify the difference between auto and manual cars, using the simple model, the difference of mpg between automatic and manual cars is calculated to be 7.24, however, the next model which takes into account "wt", "qsec" and "am" states that mpg of automatic cars is 2.94. Although the latter difference in mpg is much less than that of the simple model, it is still statistically significant.

Conclusion

Based on the above analysis, it is apparent that cars with manual transmission consume statistically significantly less fuel than that of automatic cars. However considering other influencing variables such as wt and qsec the difference is not as significant.

Appendix

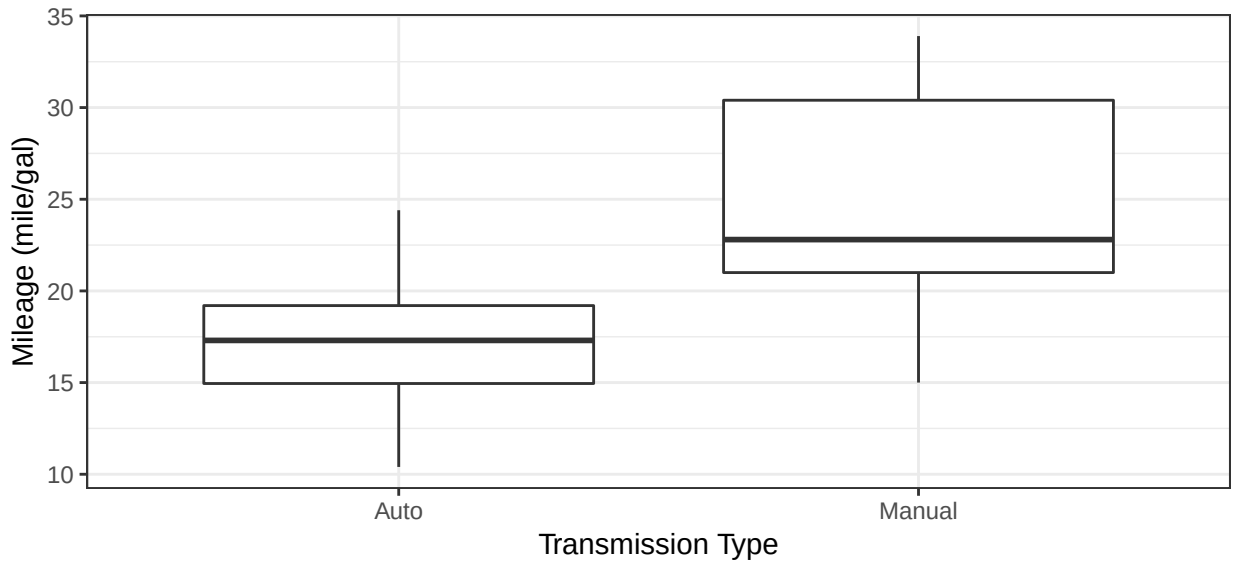

Figure 1. Mileage vs transmission type

The output of models comparisons:

```
most.eff.model <- step(lm(data = mtcars, mpg ~ wt+cyl+qsec+am+hp), direction = "both")
```

```
## Start:  AIC=63.47
## mpg ~ wt + cyl + qsec + am + hp
##
##        Df Sum of Sq    RSS    AIC
## - cyl   1     0.249 160.07 61.515
## - hp    1     7.967 167.78 63.022
## - qsec  1    10.180 170.00 63.442
## <none>              159.82 63.465
## - am    1    15.402 175.22 64.410
## - wt    1    60.723 220.54 71.771
##
## Step:  AIC=61.52
## mpg ~ wt + qsec + am + hp
##
##        Df Sum of Sq    RSS    AIC
## - hp    1     9.219 169.29 61.307
## <none>              160.07 61.515
## - qsec  1    20.225 180.29 63.323
## + cyl   1     0.249 159.82 63.465
## - am    1    25.993 186.06 64.331
## - wt    1    78.494 238.56 72.284
##
## Step:  AIC=61.31
## mpg ~ wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
```

```
## <none>              169.29 61.307
## + hp     1     9.219 160.07 61.515
## + cyl    1     1.501 167.78 63.022
## - am     1    26.178 195.46 63.908
## - qsec   1   109.034 278.32 75.217
## - wt     1   183.347 352.63 82.790
```

```
par(mfrow=c(2,2))
plot(lm(mpg~am,data = mtcars))
```

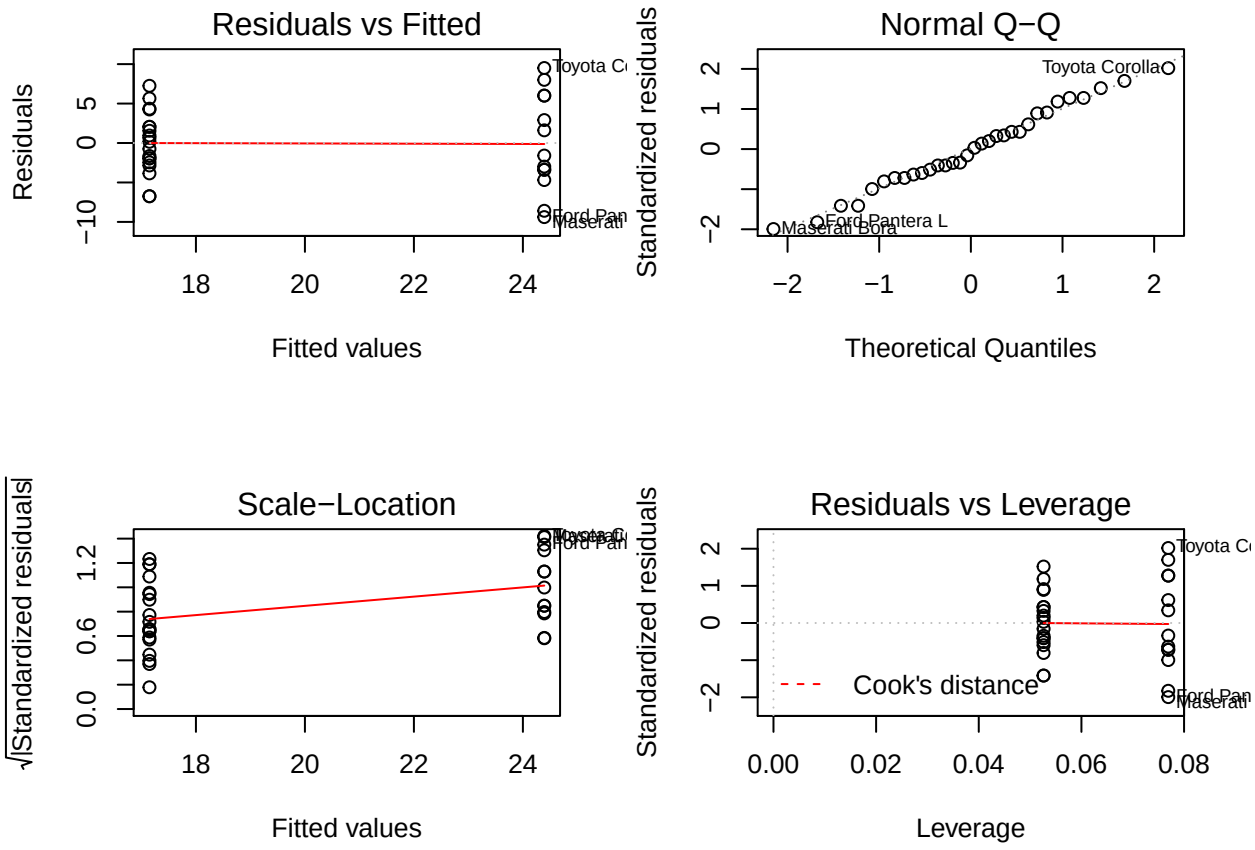

Figure 2. model 1 residuals

```
par(mfrow=c(2,2))
plot(most.eff.model)
```

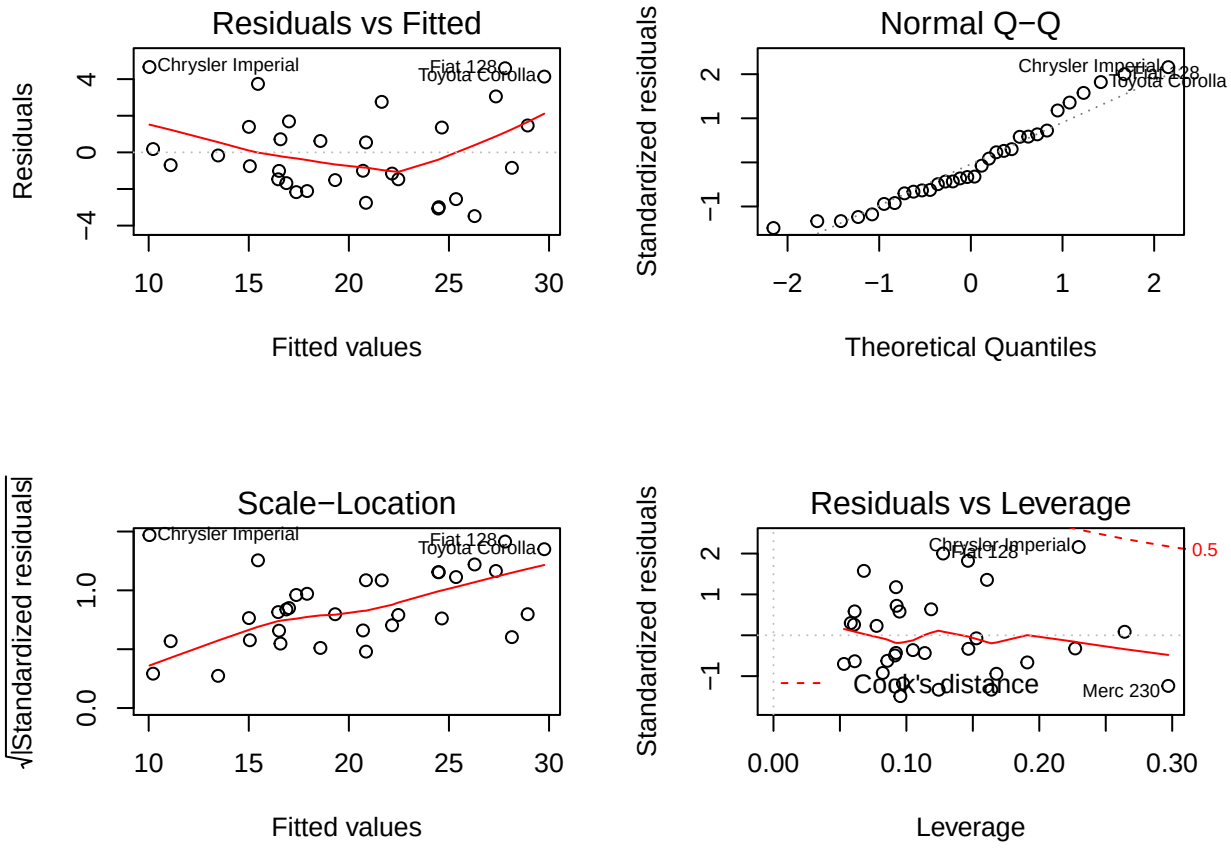

Figure 3. The most efficient model residuals

The Rmd file to generate this report can be found here.