Exploring the Impact of Certain Vehicle Characteristics on Vehicle Miles per Gallon (MPG)

Andrea Vallebueno

February 3rd, 2019

Executive summary

In the following article, we seek to explore the relationship between miles per gallon (MPG) and a set of variables including number of cylinders, displacement, gross horsepower, and others from the mtcars dataset. In particular, we seek to understand the relationship between MPG and transmission (automatic or manual). After conducting some exploratory analysis, we describe a model selection strategy and fit a series of models to examine and quantify this relationship. We find that the mean MPG is higher for cars with a manual transmission, and is lower for cars with more cylinders or more horsepower.

Setting up the environment

Let us begin by loading the necessary libraries.

```
library(ggplot2)
library(GGally)
```

Exploratory Analysis

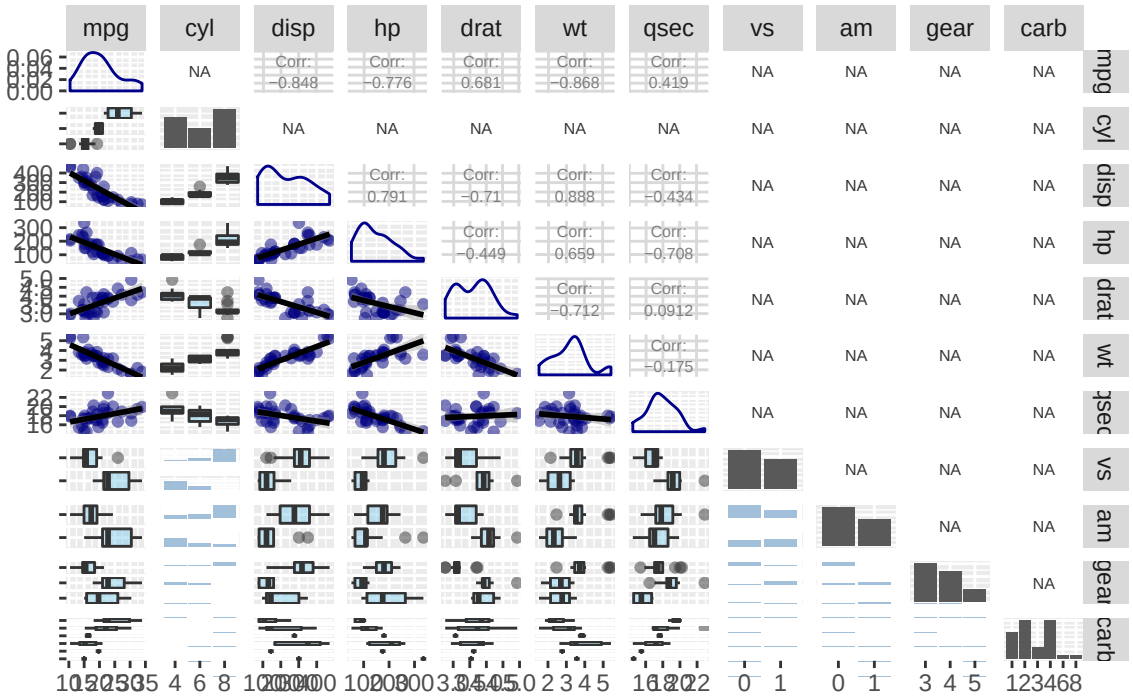
We begin by downloading the 1974 Motor Trend US magazine data set mtcars, which as per R Documentation "comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles." The 11 numeric variables include Miles/(US) gallon (mpg), number of cylinders (cyl), displacement (cu.in.) (disp), gross horsepower (hp), rear axle ratio (drat), weight (1000 lbs) (wt), 1/4 mile time (qsec), engine (V-shaped / straight) (vs), transmission (automatic / manual) (am), number of forward gears (gear) and number of carburetors (carb).

```
data(mtcars)
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

We now convert the number of cylinders, engine, transmission, number of forward gears and number of carburetors variables into factor variables as they are currently numeric, in order to correctly analyze these.

Figure 1: Relationship between all variables



From Figure 1, we observe positive relationships between MPG and the number of cylinders, rear axle ratio, number of seconds (qsec) and the number of carburetors. On the other hand, we note negative relationships between MPG and displacement, horsepower, weight, engine and transmission. An important note for our model selection strategy is that several of the explanatory variables that we will be including as regressors are correlated (e.g. displacement and horsepower, rear axle ratio and weight). We know that this can result in bias in the coefficients of interest in case we omit important variables.

Models and model selection

For our model selection strategy, we shall begin with the simplest model, that which explains MPG with the single regressor am or transmission type. We will then proceed to perform a series of nested likelihood ratio tests to determine the importance of adding additional variables.

Model 1

We begin by explaining MPG based solely on transmission type.

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

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am1            7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

```
model1beta0 <- coefficients(model1)[1]
model1beta1 <- coefficients(model1)[2]
```

With the previous linear model, we find estimates of 17.1473684 and 7.2449393 for Beta0 and Beta1, respectively, which mean that the mean MPG for an automatic transmission is 17.1473684, while the change in mean in MPG from automatic to a manual transmission is positive at 7.2449393.

Models with the addition of new variables

We will now test the relationship found between MPG and transmission type in Model 1 through the addition of several variables. Each model adds an additional variable all the way up to Model 10, which includes all 10 explanatory variables.

```
model2 <- update(model1, mpg ~ am + cyl + disp)
model3 <- update(model2, mpg ~ am + cyl + disp)
model4 <- update(model3, mpg ~ am + cyl + disp + hp)
model5 <- update(model4, mpg ~ am + cyl + disp + hp + drat)
model6 <- update(model5, mpg ~ am + cyl + disp + hp + drat + wt)
model7 <- update(model6, mpg ~ am + cyl + disp + hp + drat + wt + qsec)
model8 <- update(model7, mpg ~ am + cyl + disp + hp + drat + wt + qsec + vs)
model9 <- update(model8, mpg ~ am + cyl + disp + hp + drat + wt + qsec + vs + gear)
model10 <- update(model9, mpg ~ am + cyl + disp + hp + drat + wt + qsec + vs + gear + carb)
anova(model1, model2, model3, model4, model5, model6, model7, model8, model9, model10)
```

```
## Analysis of Variance Table
##
## Model  1: mpg ~ am
## Model  2: mpg ~ am + cyl + disp
## Model  3: mpg ~ am + cyl + disp
## Model  4: mpg ~ am + cyl + disp + hp
## Model  5: mpg ~ am + cyl + disp + hp + drat
## Model  6: mpg ~ am + cyl + disp + hp + drat + wt
## Model  7: mpg ~ am + cyl + disp + hp + drat + wt + qsec
## Model  8: mpg ~ am + cyl + disp + hp + drat + wt + qsec + vs
## Model  9: mpg ~ am + cyl + disp + hp + drat + wt + qsec + vs + gear
## Model 10: mpg ~ am + cyl + disp + hp + drat + wt + qsec + vs + gear + carb
##    Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1      30 720.90
## 2      27 230.46  3    490.44 20.3665 1.512e-05 ***
## 3      27 230.46  0      0.00
## 4      26 183.04  1     47.42  5.9078   0.02809 *
## 5      25 182.38  1      0.66  0.0820   0.77855
```

```
## 6      24 150.10  1     32.28  4.0216   0.06331 .
## 7      23 141.21  1      8.89  1.1081   0.30916
## 8      22 139.02  1      2.18  0.2719   0.60964
## 9      20 134.00  2      5.02  0.3128   0.73606
## 10     15 120.40  5     13.60  0.3388   0.88144
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the Analysis of Variance Table, we note that only the variables number of cylinders and horsepower appear to be necessary inclusions using a significance level of 5%. We note that these two variables show significant correlation (see Figure 1) with other variables, meaning that they most likely capture significant explanatory power.

With this in mind, let us look at the following model including only number of cylinders and horsepower.

```
model <- lm(mpg ~ am + cyl + hp, data = mtcars)
summary(model)
```

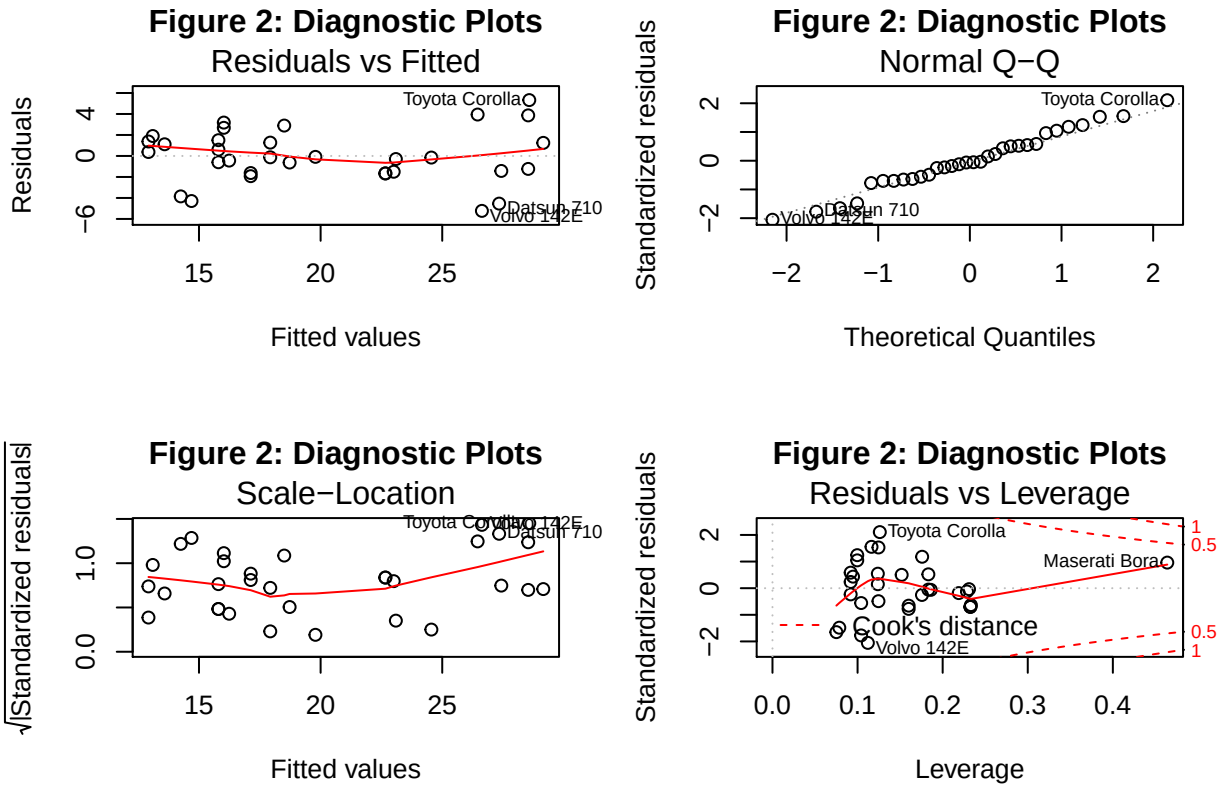
```
##
## Call:
## lm(formula = mpg ~ am + cyl + hp, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.231 -1.535 -0.141  1.408  5.322
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 27.29590    1.42394  19.169  < 2e-16 ***
## am1          4.15786    1.25655   3.309  0.00266 **
## cyl6        -3.92458    1.53751  -2.553  0.01666 *
## cyl8        -3.53341    2.50279  -1.412  0.16943
## hp          -0.04424    0.01458  -3.035  0.00527 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.703 on 27 degrees of freedom
## Multiple R-squared:  0.8249, Adjusted R-squared:  0.7989
## F-statistic: 31.79 on 4 and 27 DF,  p-value: 7.401e-10
```

```
modelbeta0 <- coefficients(model)[1]
modelbeta1 <- coefficients(model)[2]
modelbeta2 <- coefficients(model)[3]
modelbeta3 <- coefficients(model)[4]
modelbeta4 <- coefficients(model)[5]
```

From the coefficient estimates, we find that holding number of cylinders and horsepower constant, the transmission type appears to have a larger impact on MPG than if these two variables were disregarded. Interpreting these coefficients, we find that the mean MPG for an automatic transmission with 4 cylinders is 27.2958993, while the change in the mean MPG for an automatic transmission with 6 cylinders is -3.9245785 and with 8 cylinders is -3.5334139. On the other hand, the change in mean MPG for a manual transmission with 4 cylinders from an automatic transmission is 4.1578565, and for a manual transmission with 6 cylinders is 0.233278 and 8 cylinders is 0.6244425. Holding other variables constant, horsepower seems to have a negative relationship with MPG.

We will now perform residual plots and diagnostics to take a closer look at our model. We see that residuals correctly display a lack of any systematic pattern in figure 2.1 Residuals vs Fitted, Figure 2.3 Scale-Location and in figure 2.4 Residuals vs Leverage. Moreover, we note that viewing Figure 2.2 Normal Q-Q, our data appears to plausibly come from a Normal distribution.

```
par(mfrow = c(2, 2))
plot(model, main ="Figure 2: Diagnostic Plots")
```



Conclusion

To conclude, we have fitted a series of models seeking to explain the relationship between car characteristics and MPG. Being most interested in the impact of the transmission type, we have included this variable and performed an Analysis of Variance to include two additional variables (cyl and hp) to better explain the impact in MPG. Our estimates have found that the mean MPG is higher for cars with a manual transmission, and is lower for cars with more cylinders or more horsepower. The model that we have fitted has an R-squared of 82.5%, suggesting that the chosen covariates explain a relatively significant variation in the outcome MPG.