Automatic vs Manual Transmission Influence on MPG

Mateus Melo

12/10/2020

Executive Summary

In this analysis, we investigate whether or not there is a significant effect in the type of transmission used (automatic or manual) and the mpg value. This is done performing three steps: an exploratory analysis, where we try to understand the data and observe some trends; a statistical inference, where we try to build strong statistical evidence of the relation between the variables; the build of a regression model that best fits the data. By the end, we interpret our model to see it's limitations.

Exploratory Analysis

Let us start by loading and getting some general information about the data.

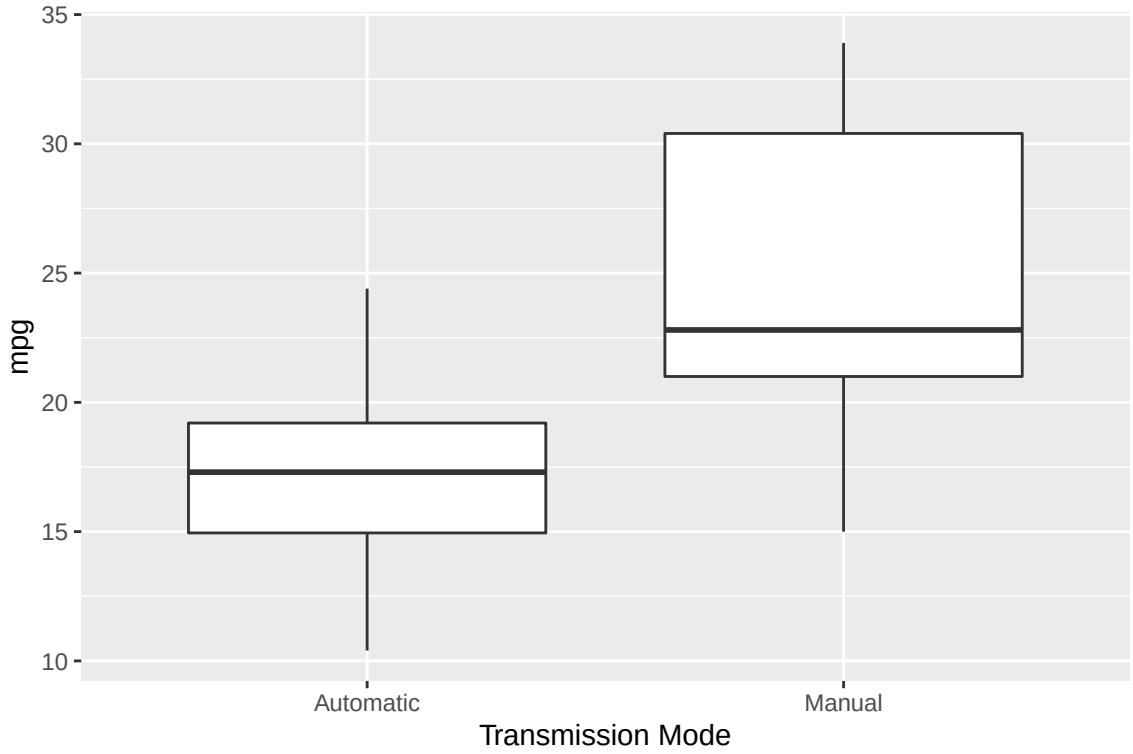
```
data("mtcars")
str(mtcars)
```

```
## '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 have 32 observations of 11 variables, all of them being of the numeric type. However, checking the documentation of the data, we conclude that the cyl, vs, am gear and carb are categorical variable. Therefore, it will be better to treat them as factors.

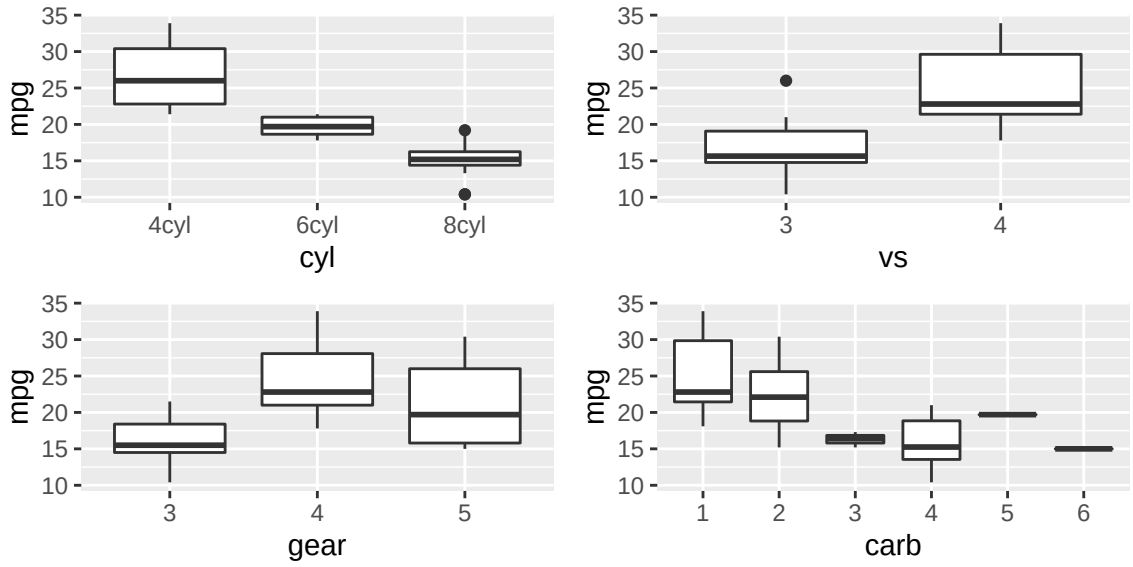
We are interested in the mpg relation with the am variable. Let's try to visualize this with a boxplot.

```
library(ggplot2)
ggplot(mtcars,aes(am,mpg))+geom_boxplot()+xlab("Transmission Mode")
```



There appears to have a difference between using an automatic or manual transmission. Let us see the mpg relation with the rest of the categorical variables.

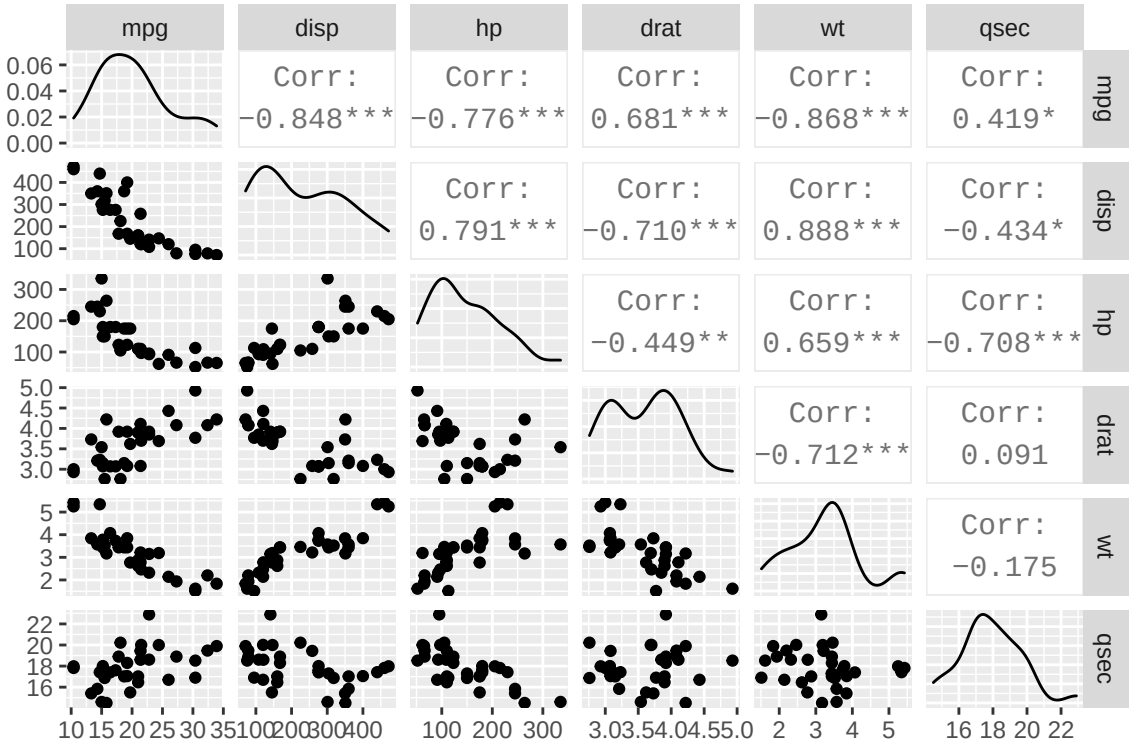
```
library(gridExtra)
p1 <- ggplot(mtcars, aes(cyl,mpg))+geom_boxplot()
p2 <- ggplot(mtcars, aes(vs,mpg))+geom_boxplot()
p3 <- ggplot(mtcars, aes(gear,mpg))+geom_boxplot()
p4 <- ggplot(mtcars, aes(carb,mpg))+geom_boxplot()
grid.arrange(p1,p2,p3,p4,nrow=2)
```



Mpg seems to be related to the cyl and vs variables.

To finish our exploratory analysis, let's take a look a the relation between the mpg and the numeric variables.

```
library(GGally)
ggpairs(mtcars,columns = c(1,3,4,5,6,7))
```



We have seen that the mpg variable is strongly correlated with all the numeric variables with the exception of qsec.

Statistical Inference

To see if the am, vs and cyl variables actually affects the mpg, we are going to perform a t test. As we have seen, the data is not paired and the variance is different.

```
am_p.value <- t.test(mpg~am, data=mtcars)$p.value
vs_p.value <- t.test(mpg~vs, data=mtcars)$p.value
cyl_4_6_p.value <- t.test(mpg~cyl,data=subset(mtcars,cyl=="4cyl" | cyl=="6cyl"))$p.value
cyl_6_8_p.value <- t.test(mpg~cyl,data=subset(mtcars,cyl=="8cyl" | cyl=="6cyl"))$p.value
cbind(am_p.value,vs_p.value,cyl_4_6_p.value,cyl_6_8_p.value)
```

```
##       am_p.value   vs_p.value cyl_4_6_p.value cyl_6_8_p.value
## [1,] 0.001373638 0.0001098368    0.0004048495    4.540355e-05
```

Since all the tests returned very low p values (all of them being below 0.01), we can assume that these variables indeed affect the mpg.

Regression Model

We are going to start to build our model including all the variables which we considered relevant in the steps above and see how well it fits the data.

```
df <- subset(mtcars, select=c("mpg", "am", "cyl","vs","disp", "hp", "drat", "wt" ))
fit1 <- lm(mpg~.,df)
summary(fit1)$coef
```

```
##                 Estimate Std. Error      t value     Pr(>|t|)
## (Intercept) 29.829969134 6.74446788  4.422879559 0.0001962074
## amManual     2.558988828 1.74302127  1.468134026 0.1556117761
## cyl6cyl     -2.055523435 1.80310789 -1.139989150 0.2660238246
## cyl8cyl     -0.023304443 3.81651017 -0.006106218 0.9951806281
## vs4          2.004897600 1.82994849  1.095603300 0.2845926673
## disp         0.004360163 0.01303611  0.334468226 0.7410571328
## hp          -0.035794756 0.01463423 -2.445960216 0.0225138202
## drat         0.388141033 1.46606024  0.264751080 0.7935593982
## wt          -2.594622674 1.20129538 -2.159854031 0.0414485707
```

Only the intercept and the hp coefficient had a p value low enough to be considered relevant. Let us see the variance inflation.

```
library(car)
vif(fit1)
```

```
##           GVIF Df GVIF^(1/(2*Df))
## am    3.780870  1        1.944446
## cyl  17.119704  2        2.034108
## vs    4.251750  1        2.061977
## disp 13.047000  1        3.612063
## hp    5.031751  1        2.243156
## drat  3.071076  1        1.752449
## wt    6.905333  1        2.627800
```

We see that the variation inflation is very large for all the variables. This is not a surprise since they are very correlated, as we have seen in the exploratory analysis. In fact, it is pretty obvious that a car with a larger disp will have a large weight. The same can be said about the amount of cyl. Also, as the disp increases, so the hp does. Let us build several models with different variables and see how they compare with each other.

```
fit2 <- lm(mpg~am+vs, df)
fit3 <- lm(mpg~am+vs+wt, df)
fit4 <- lm(mpg~am+vs+wt+hp, df)
fit5 <- lm(mpg~am+vs+disp+hp+wt, df)
fit6 <- lm(mpg~am+vs+cyl+disp+hp+wt, df)
fit7 <- lm(mpg~am+vs+cyl+disp+hp+wt+drat, df)
anova(fit2,fit3,fit4,fit5,fit6,fit7)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am + vs
## Model 2: mpg ~ am + vs + wt
```

```
## Model 3: mpg ~ am + vs + wt + hp
## Model 4: mpg ~ am + vs + disp + hp + wt
## Model 5: mpg ~ am + vs + cyl + disp + hp + wt
## Model 6: mpg ~ am + vs + cyl + disp + hp + wt + drat
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     29 353.49
## 2     28 216.32  1   137.170 22.1156 9.765e-05 ***
## 3     27 168.96  1    47.352  7.6345   0.01107 *
## 4     26 164.31  1     4.652  0.7501   0.39539
## 5     24 143.09  2    21.221  1.7107   0.20295
## 6     23 142.66  1     0.435  0.0701   0.79356
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Including wt and hp in our model provokes significant changes in the variance. Therefore, we can assume that fit4 can serve as a valid model.

Model Interpreting

Let us check some general information of the chosen model.

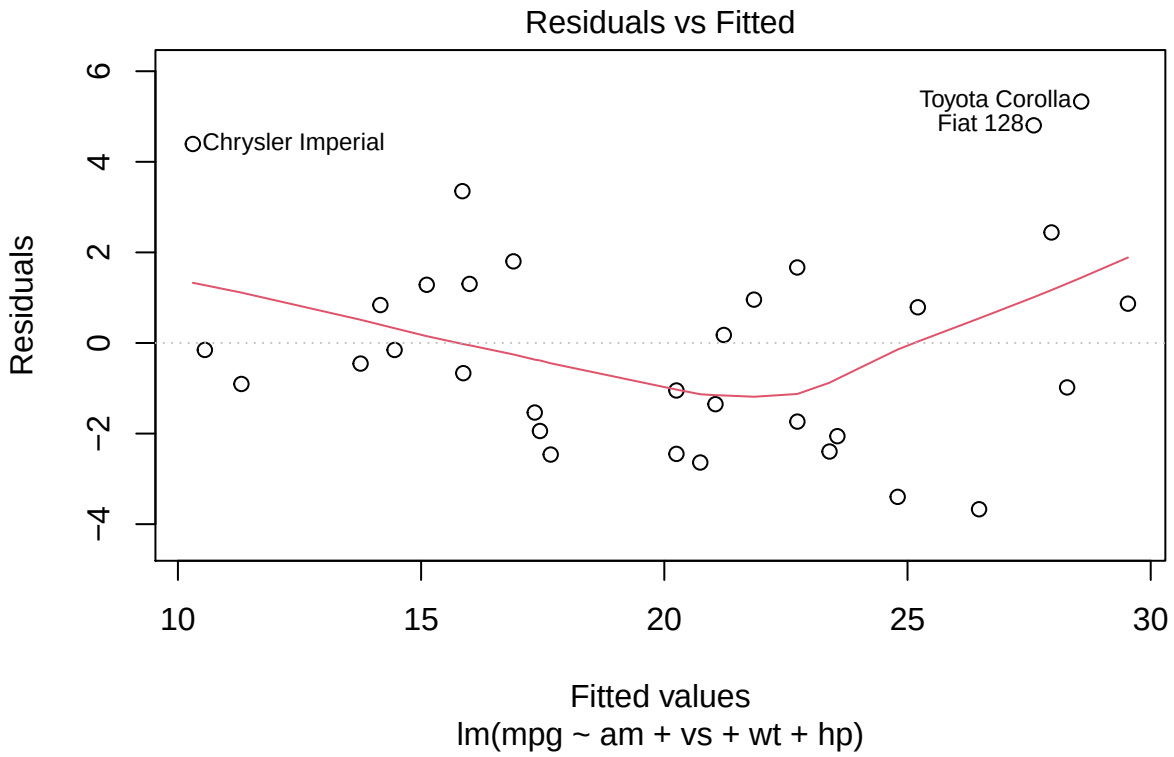
```
summary(fit4)
```

```
##
## Call:
## lm(formula = mpg ~ am + vs + wt + hp, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6710 -1.7876 -0.3044  1.2895  5.3296
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 31.07879    3.39277   9.160    9e-10 ***
## amManual     2.41714    1.37938   1.752   0.0911 .
## vs4          1.78555    1.32714   1.345   0.1897
## wt          -2.59100    0.91740  -2.824   0.0088 **
## hp          -0.03010    0.01094  -2.751   0.0105 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.502 on 27 degrees of freedom
## Multiple R-squared:  0.8499, Adjusted R-squared:  0.8277
## F-statistic: 38.23 on 4 and 27 DF,  p-value: 9.445e-11
```

We have a high R^2 value and low p value for the model. Therefore, we are able to assume that the model is relevant. The wt and hp have low p values, therefore we can assume that these variables indeed have a linear relation with the outcome. The intercept represents the expected value when the am value is set to 0 (automatic transmission), vs is set to 3 and the other variables set to 0. It also has a very low p value, so we can assume it is indeed relevant. The amManual estimate represents the mean expected change of the outcome when we have manual transmission. Since it's p value is larger than 0.05, we fail to reject the null hypothesis that there is a significant change in the outcome for different types of transmission.

To end up, let us see the residuals behavior against the fitted values.

```r
plot(fit4,which=1)
```



There is no clear pattern and the residuals variance seems to be the same across the fitted values.

Conclusion

Although there is a difference in the mpg when we change from automatic to manual transmission (the change is around 2.42), we lack of strong statistical evidence to conclude that the manual transmission is indeed better than the automatic one.