Relationship between transmission type and car performance in miles per gallon (MPG)

Nathalie Descusse-Brown

March 6, 2019

Executive Summary

The present paper investigates relationship between transmission type and car performance in miles per gallon based on some 1974 motor data. A linear regression is performed to analyse the data, which finds no statistically significant difference between mpg for automatic versus manual tranmission cars. However, further analysis is recommended based on the small sample size.

Data Sources

US automobile magazine Motor Trend decided to investigate the relationship between car performance measured in miles per gallon (mpg) and the tranmission type for data extracted from the 1974 edition of the magazine. In particular, this paper seeks to address the following questions: 1. Is an automatic or manual transmission better for MPG? 2. What is the MPG difference between automatic and manual transmissions?

The data used for this analysis comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-1974 models).

Data Processing and Exploratory Analysis

First, some exploratory analysis was performed to identify some potential trends in the data and see whether there was any indication of association between miles per gallon and transmission type.

The exploratory boxplot showing mpg per transmission type (see Appendix A1) indicated that cars with manual transmission perform better in terms of miles per gallon than cars with automatic transmission. But is it true, or are there other variables that may explain this difference in mpg?

To answer this question, a full linear regression analysis was performed. The first task was to determine which covariables to include in the regression analysis.

Intuitively assumptions were made that car weight and possibly horsepower would effect miles per gallon, so the correlation between the various parameters in the dataset was used, as seen in the matrix plot shown in Appendix A2, to try and identify which variables appear to be correlated with which. It can be seen that mpg seems to be strongly correlated with number of cylinders, displacement and weight.

Some more exploratory plot exploring the relationship between mpg, transmission type and number of cylinders further are presented in Appendix A3. It can be observed that in general a lower number of cylinders seems to be associated with higher mpg, and that cars with automatic transmission appear to typically have a higher number of cylinders.

Model Selection and Results

The exploratory plots indicated an apparent large difference in miles per gallon performance between cars with manual and automatic transmission and also some coorelation between some variables, which were explored further in Appendix B with the aim to select the covariables for the linear regression model presented in this paper. The covariables that were found to have a significant effect in our linear regression are weight, number of cylinders, and gross horsepower. The displacement, which was found to be strongly correlated to the mpg was not included as its inclusion in the model was shown to have insignificant effect, which can probably be explained by the strong correlation found in the exploratory plots between weight and displacement.

```
library(MASS)
library(car)
lmregression <- lm(mpg ~ I(factor(am)) + wt + I(factor(cyl)) + hp,data=mtcars)
summary(lmregression)$coefficients
```

```
##                    Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)     33.70832390 2.60488618 12.940421 7.733392e-13
## I(factor(am))1   1.80921138 1.39630450  1.295714 2.064597e-01
## wt              -2.49682942 0.88558779 -2.819404 9.081408e-03
## I(factor(cyl))6 -3.03134449 1.40728351 -2.154040 4.068272e-02
## I(factor(cyl))8 -2.16367532 2.28425172 -0.947214 3.522509e-01
## hp              -0.03210943 0.01369257 -2.345025 2.693461e-02
```

The results from the linear regression analysis appear to indicate that there is no statistical significance in terms of miles per gallons between cars with automatic transmission and cars with manual transmission.

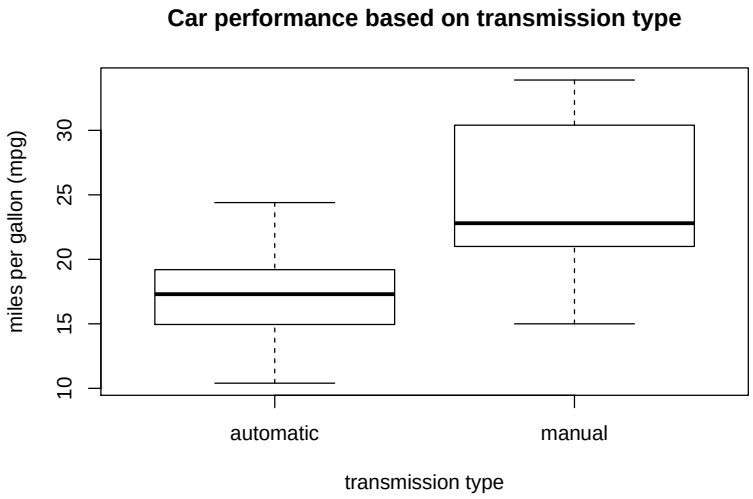
Discussion

Any linear regression model makes a number of assumptions that require checking to assess the suitability of the model. In particular the residual distribution should be checked for any remaining pattern that would indicate some additional covariables should have been included. Also, a linear regression makes the assumptions of normal distribution of the residuals and this need to be verified. The correspondong plots can be found in Appendix C.

The plot of the residuals does not seem to have any pattern to it. However it can be noticed that the lower and upper quantiles deviate slightly from a normal distribution, so other regression models may be better suited to the data. It should also be noted that the sample size is very small, which leads to query the conclusions of the analysis given the number of degrees of freedom in the model. For these reasons, it is recommended that further data are collected and that the apparent insignificance of the type of transmission on the car performance measured in miles per gallon is investigated further.

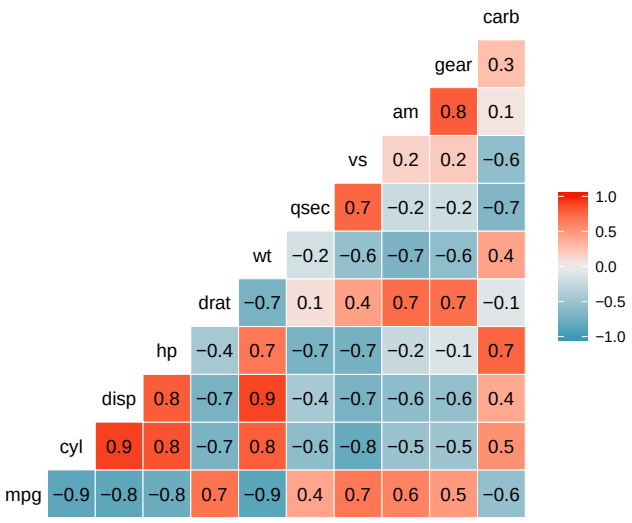
Appendix A1: Exploratory plots between mpg and transmission type

```
library(datasets)
data(mtcars)
for (i in 1:dim(mtcars)[1]) {
        if (mtcars$am[i]==1) {mtcars$trans[i] <- "manual"}
        else mtcars$trans[i] <- "automatic"
}
boxplot(mtcars$mpg ~ trans, data=mtcars, xlab="transmission type",ylab="miles per gallon (mpg)"
        ,main="Car performance based on transmission type")
```



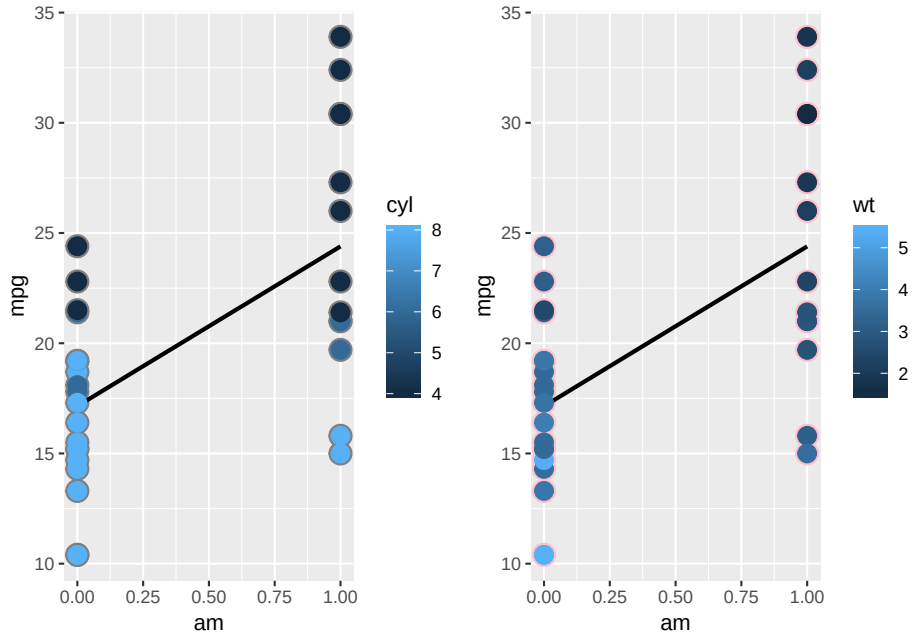
Appendix A2: Matrix plot of correlation between the dataset variables

```
library(ggplot2)
library(GGally)
ggcorr(mtcars, palette = "RdBu", label = TRUE)
```



Appendix A3: Exploratory plots on relationship between mpg, transmission type and number of cylinders

```
library(gridExtra)
g = ggplot(mtcars,aes(y=mpg,x=am,colour=cyl))+geom_point(colour="grey50",size=5)+
        geom_smooth(method=lm,se=FALSE,colour="black") +geom_point(size=4)
g2 = ggplot(mtcars,aes(y=mpg,x=am,colour=wt))+geom_point(colour="pink",size=5)+
        geom_smooth(method=lm,se=FALSE,colour="black")+geom_point(size=4)
grid.arrange(g, g2, ncol=2)
```



Appendix B: Investigation of effect of inclusion of given covariable for model fitting

```
lmregression1 <- lm(mpg ~ I(factor(am)),data=mtcars)
lmregression11 <- lm(mpg ~ I(factor(am)) + wt,data=mtcars)
lmregression21 <- lm(mpg ~ I(factor(am)) + wt + I(factor(cyl)), data=mtcars)
lmregression22 <- lm(mpg ~ I(factor(am)) + wt + hp,data=mtcars)
lmregression23 <- lm(mpg ~ I(factor(am)) + wt + disp,data=mtcars)
lmregression24 <- lm(mpg ~ I(factor(am)) + wt + drat,data=mtcars)
lmregression25 <- lm(mpg ~ I(factor(am)) + wt + qsec,data=mtcars)
lmregression26 <- lm(mpg ~ I(factor(am)) + wt + I(factor(vs)),data=mtcars)
lmregression27 <- lm(mpg ~ I(factor(am)) + wt + I(factor(gear)),data=mtcars)
lmregression28 <- lm(mpg ~ I(factor(am)) + wt + I(factor(carb)),data=mtcars)

anova(lmregression11,lmregression21)$`Pr(>F)`
```

```
## [1]          NA 0.003473216
```

```
anova(lmregression11,lmregression22)$`Pr(>F)`
```

```
## [1]           NA 0.0005464023
```

```
anova(lmregression11,lmregression23)$`Pr(>F)`
```

```
## [1]        NA 0.0678774
```

```
anova(lmregression11,lmregression24)$`Pr(>F)`
```

```
## [1]        NA 0.2849371
```

```
anova(lmregression11,lmregression25)$`Pr(>F)`
```

```
## [1]           NA 0.0002161737
```

```
anova(lmregression11,lmregression26)$`Pr(>F)`
```

```
## [1]          NA 0.008454158
```

```
anova(lmregression11,lmregression27)$`Pr(>F)`
```

```
## [1]        NA 0.1200295
```

```
anova(lmregression11,lmregression28)$`Pr(>F)`
```

```
## [1]        NA 0.2470563
```

```
lmregression31 <- lm(mpg ~ I(factor(am)) + wt + I(factor(cyl)) + hp, data=mtcars)
lmregression32 <- lm(mpg ~ I(factor(am)) + wt + I(factor(cyl)) + qsec, data=mtcars)
lmregression33 <- lm(mpg ~ I(factor(am)) + wt + I(factor(cyl)) + I(factor(vs)), data=mtcars)

anova(lmregression21,lmregression31)$`Pr(>F)`
```

```
## [1]         NA 0.02693461
```

```
anova(lmregression21,lmregression32)$`Pr(>F)`
```

```
## [1]         NA 0.06085519
```

```
anova(lmregression21,lmregression33)$`Pr(>F)`
```

```
## [1]        NA 0.5200389
```

Based on the above comparison of effects of covariables, we selected am, wt, cyl and hp as the covariable for our model. Indeed, by adding these covariable one at a time we see that they are the only covariables shown by anova to have a significant effect.

Appendix C: Linear regression models plots

```
par(mfrow = c(2,2))
plot(lmregression)
```