Motor Trend: Effect of Transmission on Fuel Consumption

Executive summary

In this report I explore the relationship between consumption in miles per gallon and transmission type (automatic vs. manual). Even though it initially looks like manual transmission is associated with less consumption, by analysing counfounding variables, I find that, for a given number of cylinders and weight, this difference is not significant. As such, no clear evidence was found in favour of either automatic or manual transmission.

1. Exploratory data analysis

- I'm particularly interested in understanding the relationship between the consumption **mpg** and the transmission **am**. A boxplot is a good starting point to graphically explore any potential relationship between these two variables (*Figure 1*). It seems that manual transmission is assossiated with less consumption (*as in more miles made per the same amount of fuel*).
- However, a number of other counfounding variables could be influencing this result. For example, it would seem logical to think that number of cylinders **cyl** could also influence the consumption. The previous boxplot was conditioned to this variable in order to check if the manual transmission is associated with better consumption even when the number of cylinders is hold constant (*Figure 2*). It seems that the major difference is specific for 4 cylinder cars, the consumption being more similar for both 6 and 8 cylinder. As such, the number of cylinders is a important variable to account for.

2. Model selection

Because of the existence of counfounding variables, it is important to explore the effects of including these variables in the model. I started by exploring the correlation of each variable with the consumption **mpg**.

List of highly correlated (>0.8) variables:

```
##       cyl disp    hp  drat   wt  qsec    vs  gear  carb
## [1,] TRUE TRUE FALSE FALSE TRUE FALSE FALSE FALSE FALSE
```

From the correlations it's clear that the variables more correlated with consumption are the number of cylinders **cyl**, the displacement **disp** and the weight **wt**. From this information, 5 models were chosen to be explored:

```
fit1<-lm(mpg~am,mtcars)
fit2<-lm(mpg~am+cyl,mtcars)
fit3<-lm(mpg~am+cyl+wt,mtcars)
fit4<-lm(mpg~am+cyl+wt+disp,mtcars)
fit5<-lm(mpg~.,mtcars)
```

The variance was studied for each of these models in order to select which is more relevant. This was done by *vif()* (car package) and *anova()*

```
##    am   cyl
## 1.376 1.376
```

```
##    am   cyl    wt
## 1.925 2.584 3.609
```

```
##    am   cyl    wt  disp
## 1.932 5.414 5.987 9.958
```

According to the variance inflation factors obtained, it would seem that adding **disp** variable to the model has a great impact on the variance of **cyl** and **wt**, suggesting multicollinearity.

```
##   Res.Df RSS Df  Pr(>F)
## 1     30 721
## 2     29 271  1 8.2e-08 ***
## 3     28 191  1  0.0028 **
## 4     27 188  1  0.5478
## 5     21 147  6  0.4684
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interestingly, the analysis of variance suggests that the addition of **cyl** and **wt** are the only significant regressor additions (alpha < 0.01). This would be in agreement with the VIF analysis, which suggested that **disp** correlated with both **cyl** and **wt**, thus being a redundant variable for the model.

The chosen model is **mpg = b0 + b1.am + b2.cyl + b3.wt**

3. Residual diagnostics

The influence of each observation on the coefficients was taken by *dfbetas()*. The observations with higher influence on b0 (expected value of mpg for automatic transmission) and b1 (difference between manual and automatic transmission) are shown:

```
##                   (Intercept)      am     cyl      wt
## Toyota Corona         -0.8296  0.6742  0.3405  0.26500
## Chrysler Imperial     -0.6178  0.3954 -0.4496  0.94704
## Ford Pantera L         0.2124 -0.2358 -0.2498  0.05010
## Maserati Bora          0.2078 -0.2248 -0.1261 -0.05209
## Fiat 128               0.2069  0.3226 -0.3268  0.12006
```

To confirm the validity of the model assumptions the residuals of the model were also plotted for analysis (*Figure 3*)

3. Quantify the uncertainty and inference

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  39.4179     2.6415 14.9228 7.425e-15
## am            0.1765     1.3045  0.1353 8.933e-01
## cyl          -1.5102     0.4223 -3.5764 1.292e-03
## wt           -3.1251     0.9109 -3.4309 1.886e-03
```

Conclusions

The mean consumption of **a car with automatic transmission is 39.4 miles/(US)gallon** (standard error 2.64), while **for manual transmission this value is increased 0.17 miles/(US)gallon** (standard error 1.30), when *both number of cylinder and weigth are taken as counfounding variables.* However, the p-value for this difference is 0.89, which means that **there is no significant difference** between the consumption of cars with automatic or manual transmission. The corresponding confidence interval for b1 is -2.4956, 2.8485, suggesting the difference between automatic and manual trnasmission is not significantly different from 0.

Annexes

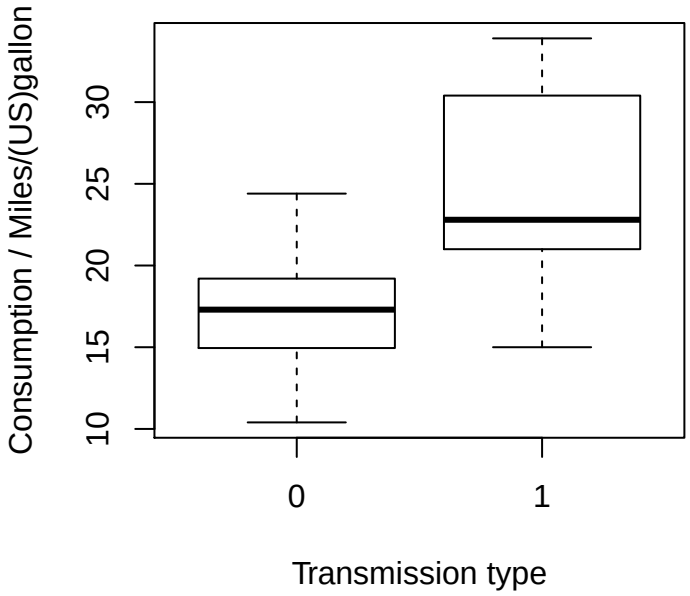


Figure 1: Representation of the consumption for cars with automatic transmission (0) or manual transmission (1)

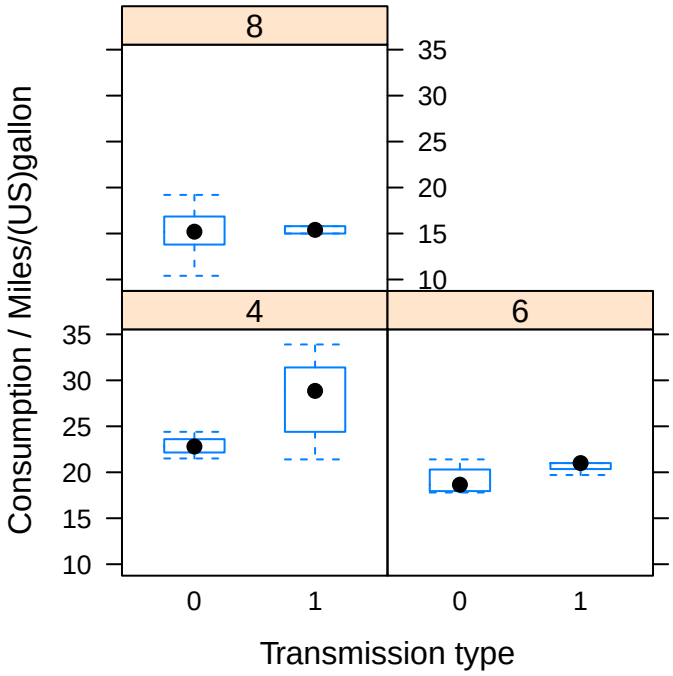
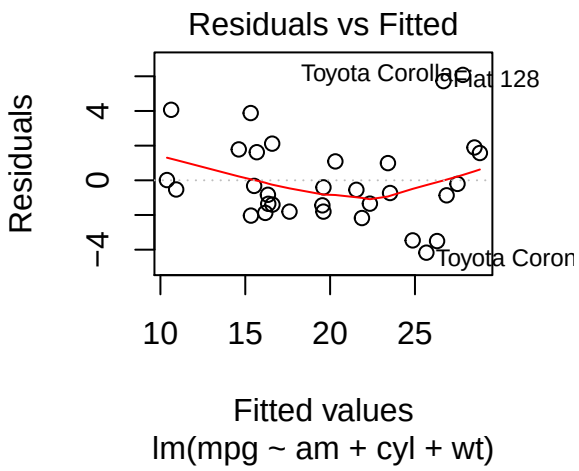
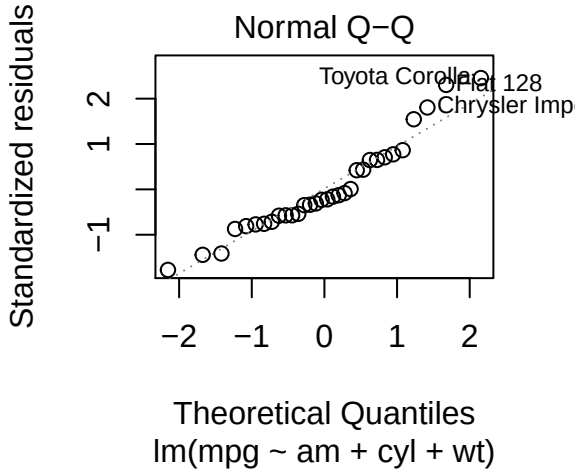


Figure 2: Representation of the consumption for cars with automatic transmission (0) or manual transmission (1) conditioned by the number of cylinders (4,6 or 8)





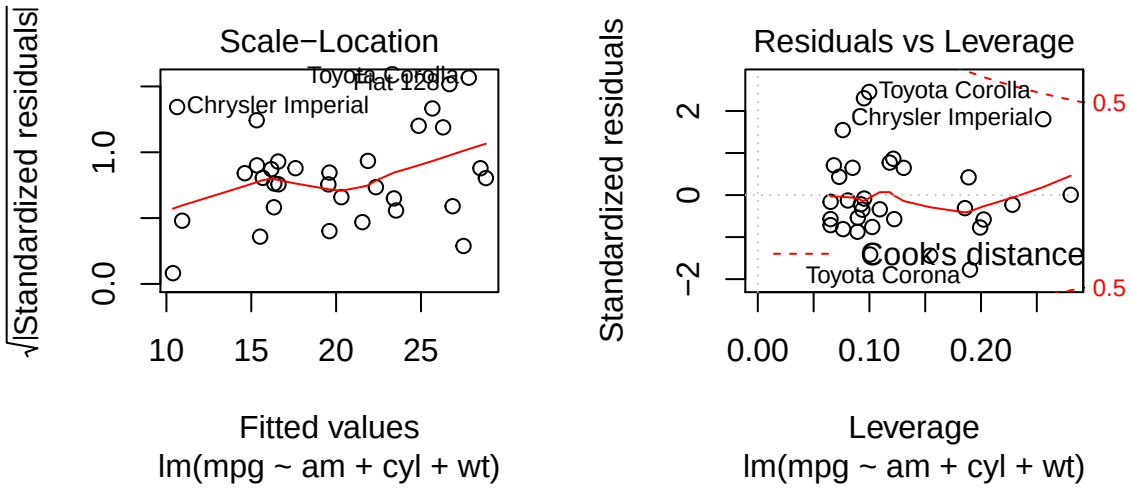


Figure 3: Residual plot, normal QQ residual plot, residual scale-location plot and Cook's distance plot