“Is an automatic or manual transmission better for car MPG ?”

by ***DT***

1. Summary

Motor Trend magazine is interested to find out the relationship between a set of variables and MPG (outcome). Basically, the company is interested to answer two questions: 1) *“Is an automatic or manual transmission better for MPG”* and 2) *“Quantify the MPG difference between automatic and manual transmissions”*. By using the `mtcars` data set we are going to answer those questions by making an exploratory data analyses and proposing various regression models between the data set variables. Residual and diagnostic analysis are made to help to choose the best regression model.

2. Exploratory data analysis

We start by loading some important libraries and data set, and also taking a look into the data

```
library(ggplot2); library(GGally); require(datasets); data(mtcars); head(mtcars,3);
##                mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4     21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag 21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710    22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
class(mtcars); dim(mtcars)
## [1] "data.frame"
## [1] 32 11
```

2.1. About *mtcars* data set:

The data set `mtcars` is a data frame with 32 observations on 11 variables. The variables definition can be found by running `?mtcars` and it is reproduced in Table 1 (see Appendix). Note that “vs” and “am” are binary variables, and three are also categorical variables, e.g., “cyl”, “gear”, and “carb”.

2.2. About the variables:

We can verify correlations between “mpg” (outcome) against the other variables (predictors) and also against the variables theimselves by making a pair wise correlation. An exploratory plot in Fig. 1 (see Appendix) is built, where it is color code by the type of transmission. The two most important features in the plot are the correlation between “mpg” and other variables shown by the linear model (first row, left to right) and the correlation values for those linear models(first column, top to bottom). It is important to note that the results in the Fig. 1 are unadjusted, i.e., paired linear relationships and not a multivariable regression. We can now make a mutivariable model considering “mpg” as outcome and the other variables as predictors.

2.3. Linear Models

The first and simplest model (lm0) is to consider that “mpg” is only depending on “am”, that is, a simple linear regression. The result is in the plot in Fig. 1 (the 3rd one in the first row from the right side). So, a positive increase in “mpg” is observed from automatic to manual transmission. However, if we look at the correlation between both (the 3rd one in the first column from the bottom) is not so strong (ca. 0.6). Also, calculating the R^2 ($= 0.3598$) means that only 36% of the variation is explained by this model. At the other extreme, we can consider a model with *all* variables, by calling lm_all model (see in Appendix). A quick interpretation of the coefficients is made by taking “cyl” as example. Its coefficient is ca. -2.65. This means that if the other predictors are kept constant, for **every unit increase** in “cyl” (in this case from 4 to 6) there will be a **decrease** in “mpg” by 2.65 units. The intercept is interpreted by being “mpg” value when all the predictions are null or at the lowest level for the factors. This may not make sense for some of variables, for instance, there is no “zero” value for “hp”, or for “gear”, and so on. This situation is avoided if the model in centered to the mean values.Therefore, the intercept is now related to the mean for those variables that were centered. Although, the R^2 value is 0.8931 and is higher than the first model (extra 9 variables), the test hypothesis of relationship between the predictors and miles per gallon failed to reject the *null* hypothesis. The *null* hypothesis is that all coefficients equal zero. All P-values are higher than 0.05 and hence for this model that consider all the predictors we failed to reject the *null* hypothesis (see Appendix for the values). “Simpson’s” paradox can be observed,for instance, on the coefficients obtained in the lm_all model, the signal of displacement variable (“disp”) is positive, but in Fig 1. the slope is negative against “mpg”. The same is observed for some other variables.

3. Linear Models - Model selection

Diferent models are tested and the best one is considered here if it is able to answer the proposed questions, and It is robust and simple (small number of variables). The strategy of model selection applied here is as follows: (1) Starting with the model with full predictors (lm_all); (2) Using function `vif` to calculate variance inflation factor of the predictors; (3) Creating a new model without the predictor with the highest *VIF* on the previous model; (4) Repeating tasks (2) and (3) until the minimum number of predictors is achieved(i.e., two), or until “am” predictor is prevalent. (5) Perfoming ANOVA on the nested models. The code is in the appendix (unfortunatelly no space to show the results for all the tested models). Note that in all tested models positive values for “am” coefficient were found. However the t-test of coefficients of “am” predictor showed that we failed to the reject *null* hypothesis, **except for the linear models lm6, lm7** and **lm8**. We can perfom a nested model analysis using ANOVA, starting with the simplest model (lm0), then lm8, lm7 and lm6 (see Appendix). From the results, the extra addition of predictor for model lm8 (Model 2 in the result), i.e., “vs” make P-value above of $\alpha = 0.05$. Therefore, lm8 with “am” and “hp” as predictors presented as the best model.

4. Residual plot and diagnostic

The diagnostic analysis is made by using `dfbetas` and `hatvalues` functions on the choosen model. Some of car models were selected based also on Residual plot (Fig.2 in the Appendix), and the `dfbetas` and `hatvalues` calculated.

```
res<-cbind(round(dfbetas(lm8)[,2:3],3),hatvalues=round(hatvalues(lm8),3))[c(18:21,28:31),]
colnames(res)<- c('dfbeta.hp','dfbeta.am','hatvalues'); res
##                dfbeta.hp dfbeta.am hatvalues
## Fiat 128          -0.287     0.292     0.104
## Honda Civic       -0.125     0.097     0.118
## Toyota Corolla    -0.403     0.402     0.105
## Toyota Corona     -0.039    -0.042     0.082
## Lotus Europa      -0.076     0.402     0.078
## Ford Pantera L    -0.082    -0.066     0.214
## Ferrari Dino      -0.091    -0.167     0.094
## Maserati Bora      0.922     0.564     0.393
```

Note that Maserati Bora has high influence on coefficients and high leverage values than the other cars. Toyota Corolla also has a high influence but relatively low leverage. Lotus has high influence only in the second coefficient (“am”) but low leverage. Similar results are also observed in the Residuals plot. Furthermore, the residual Q-Q plot(top rigth, Fig.2) shows a normality of the errors from lm8 model where the points fall over a diagonal line.

5. Conclusions

Model the coeficients of model lm8 is shown below:

```
round(summary(lm8)$coef,5)
##                  Estimate Std. Error  t value Pr(>|t|)
## (Intercept)      17.94681    0.67588 26.55307    0e+00
## I(hp - mean(hp)) -0.05889    0.00786 -7.49519    0e+00
## factor(am)1       5.27709    1.07954  4.88827    3e-05
```

All the coefficients are significant at α= 0.05. So, we can reject the *null* hypothesis. The confidence intervals can be also determined, at the same α level.

```
sumCoef <- summary(lm8)$coef
intercept <- round((sumCoef[1,1]+c(-1,1)*qt(.975,df=lm8$df)*sumCoef[1,2]),2)
hp.slope <- round((sumCoef[2,1]+c(-1,1)*qt(.975,df=lm8$df)*sumCoef[2,2]),2)
am.slope <- round((sumCoef[3,1]+c(-1,1)*qt(.975,df=lm8$df)*sumCoef[3,2]),2)
intercept
## [1] 16.56 19.33
hp.slope
## [1] -0.07 -0.04
am.slope
## [1] 3.07 7.48
```

Assuming that the best model is linear with additive independent and indentical distributed errors and also by treating `mtcars` dataset as a population, we found that manual transmission is better than automatic. Therefore, if “hp” is kept constant, the change from automatic to manual transmission increases “mpg”" by a factor of *ca.* (5.3 +/- 2.2) the value estimated for the intercept change (“mpg”) from automatic transmission to manual transmission.

Appendix

Table 1: mtcars variables definition

Variable	Definition
mpg	Miles/(US) gallon
cyl	Number of cylinders
disp	Displacement (cu.in.)
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (lb/1000)
qsec	1/4 mile time
vs	V/S (0 means a V-engine, and 1 straight engine)
am	Transmission (0 = automatic, 1 = manual)
gear	Number of forward gears
carb	Number of carburetors

Linear Models

Linear model lm0 : Only "am"

```
lm0 <-lm(mpg ~ factor(am), data=mtcars); round(summary(lm0)$coef,3)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247        0
## factor(am)1    7.245      1.764   4.106        0
```

Linear model lm_all : All predictors

```
round(summary(lm_all)$coef,3)
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            17.984      5.324   3.378    0.004
## factor(cyl)6           -2.649      3.041  -0.871    0.397
## factor(cyl)8           -0.336      7.160  -0.047    0.963
## I(disp - mean(disp))    0.036      0.032   1.114    0.283
## I(hp - mean(hp))       -0.071      0.039  -1.788    0.094
## I(wt - mean(wt))       -4.530      2.539  -1.784    0.095
## I(drat - mean(drat))    1.183      2.483   0.476    0.641
## I(qsec - mean(qsec))    0.368      0.935   0.393    0.700
## factor(vs)1             1.931      2.871   0.672    0.512
## factor(am)1             1.212      3.214   0.377    0.711
## factor(gear)4           1.114      3.800   0.293    0.773
## factor(gear)5           2.528      3.736   0.677    0.509
## factor(carb)2          -0.979      2.318  -0.423    0.679
## factor(carb)3           3.000      4.294   0.699    0.495
## factor(carb)4           1.091      4.450   0.245    0.810
## factor(carb)6           4.478      6.384   0.701    0.494
## factor(carb)8           7.250      8.361   0.867    0.399
```

Variance Inflation factor for lm_all

```
require(car)            # if you don't have the package run: install.packages("car")
round(vif(lm_all),3)
##                         GVIF Df GVIF^(1/(2*Df))
## factor(cyl)          128.121  2           3.364
## I(disp - mean(disp))  60.366  1           7.770
## I(hp - mean(hp))      28.220  1           5.312
## I(wt - mean(wt))      23.831  1           4.882
## I(drat - mean(drat))   6.810  1           2.610
## I(qsec - mean(qsec))  10.790  1           3.285
## factor(vs)             8.088  1           2.844
## factor(am)             9.930  1           3.151
## factor(gear)          50.852  2           2.670
## factor(carb)         503.212  5           1.863
```

- Linear Model lm1 : Eliminating the highest vif value, in this case it is "carb"
- Linear Model lm2 : Eliminating the highest vif value, in this case it is "cyl"

- Linear Model lm3 : Eliminating the highest vif value, in this case it is “gear”
- Linear Model lm4 : Eliminating the highest vif value, in this case it is “disp”
- Linear Model lm5 : Eliminating the highest vif value, in this case it is “qsec”
- Linear Model lm6 : Eliminating the highest vif value, in this case it is “wt”
- Linear Model lm7 : Eliminating the highest vif value, in this case it is “drat”
- Linear Model lm8 : Eliminating the highest vif value, in this case it is “vs”

ANOVA for 4 nested models

```
round(anova(lm0, lm8, lm7, lm6), 3)
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ I(hp - mean(hp)) + factor(am)
## Model 3: mpg ~ I(hp - mean(hp)) + factor(vs) + factor(am)
## Model 4: mpg ~ I(hp - mean(hp)) + I(drat - mean(drat)) + factor(vs) +
##     factor(am)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     30 720.90
## 2     29 245.44  1    475.46 60.105 <2e-16 ***
## 3     28 218.88  1     26.56  3.358  0.078 .
## 4     27 213.58  1      5.30  0.670  0.420
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

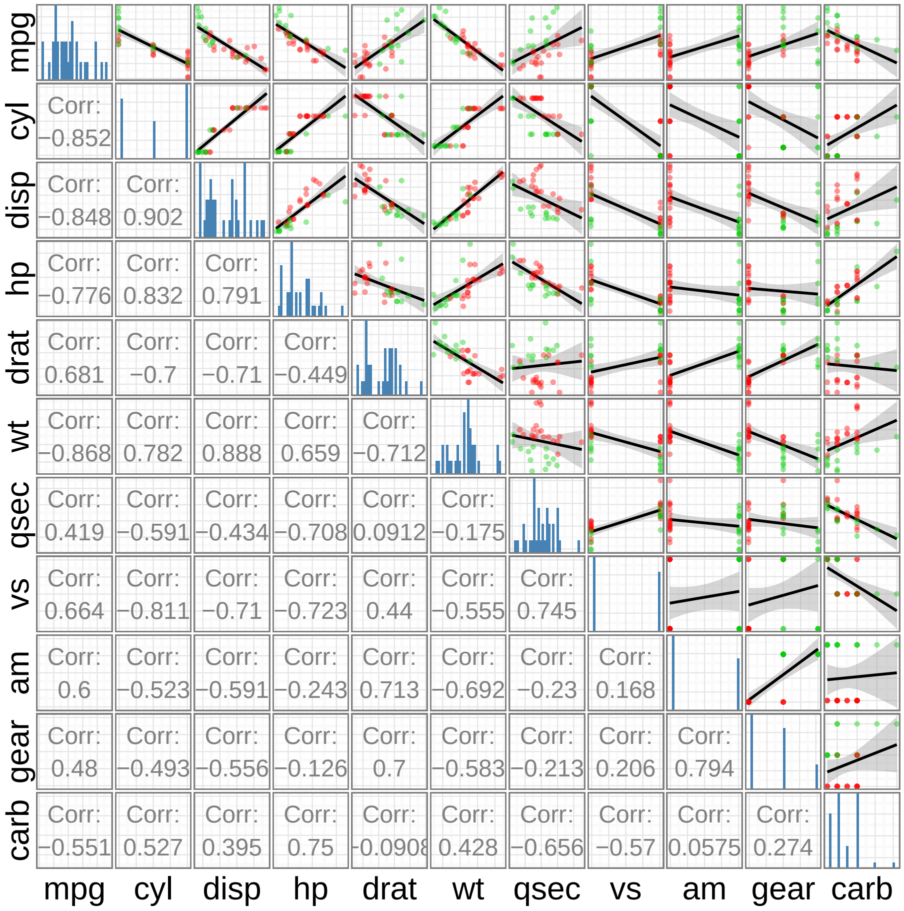

Figure 1: *Plot of the correlations of mtcars variables. Red color for automatic and Green color for manual*

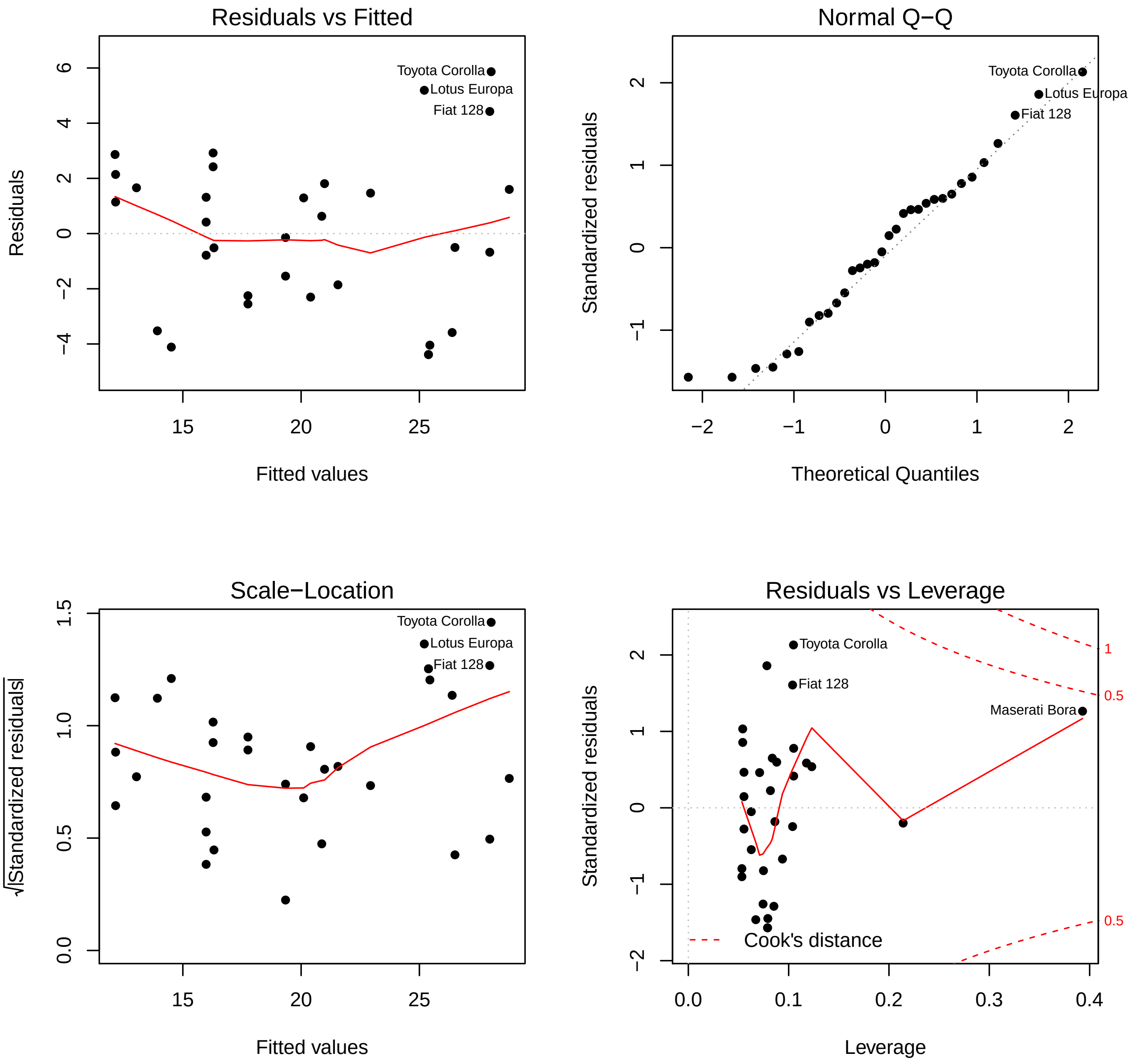

Figure 2: *Residuals analysis Plot of mtcars data of linear model 8.*