JH - M07 - Regression Project

S. Hostiou

21/10/2020

Summary

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

- Is an automatic or manual transmission better for MPG
- Quantify the MPG difference between automatic and manual transmissions

This work is part of the Regression Models course provided by Johns Hopkins University at Coursera.The “mtcars” data set is used for the purpose of this exercise.

The approach followed to answer our 2 questions will be to build a linear model with the mpg variable as a response variable and a set of regressors as inputs.
This model will provide us elements of interpretation to understand how the different regressors kept in our model can influence our miles per gallon concern.

Checking Data Set Content

Let’s start by importing the data into the R environment and by having a look at the content of the dataframe.

```
data("mtcars")
head(mtcars)
```

```
##                   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
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

The mpg (miles per gallon) variable will be our response variable. Our approach will be to configure a linear model in order to identify how the other variables can influence fuel consumption.
The following variables will be consisered as numeric continuous : disp (displacement cu in), hp (horse power),drat (rear axle ratio), wt (weight), qsec (1/4 mile time).
The remaining variables will be handled as factor variables : cyl (nb of cyl), vs (engine shape), am (auto/man transmission), gear (nb gears), carb (nb of carb).

```
# for convenience, We will work with a copy of the original data set
my_mtcars <- mtcars
```

First assumption

The first exploratory plots given in appendix 1 gives us some primary elements to answer the question. Box plot & distribution diagrams of mpg vs transmission mode (auto / man) tend to show lower mpg with automatic transmission and higher mpg with manual transmission.

Collinearity Identification

As a first step we will identify if there are strong evidences of high correlations between what would be our input variables.

```
mtcars %>% cor()
```

```
##             mpg        cyl       disp         hp        drat         wt
## mpg   1.0000000 -0.8521620 -0.8475514 -0.7761684  0.68117191 -0.8676594
## cyl  -0.8521620  1.0000000  0.9020329  0.8324475 -0.69993811  0.7824958
## disp -0.8475514  0.9020329  1.0000000  0.7909486 -0.71021393  0.8879799
## hp   -0.7761684  0.8324475  0.7909486  1.0000000 -0.44875912  0.6587479
## drat  0.6811719 -0.6999381 -0.7102139 -0.4487591  1.00000000 -0.7124406
## wt   -0.8676594  0.7824958  0.8879799  0.6587479 -0.71244065  1.0000000
## qsec  0.4186840 -0.5912421 -0.4336979 -0.7082234  0.09120476 -0.1747159
## vs    0.6640389 -0.8108118 -0.7104159 -0.7230967  0.44027846 -0.5549157
## am    0.5998324 -0.5226070 -0.5912270 -0.2432043  0.71271113 -0.6924953
## gear  0.4802848 -0.4926866 -0.5555692 -0.1257043  0.69961013 -0.5832870
## carb -0.5509251  0.5269883  0.3949769  0.7498125 -0.09078980  0.4276059
##             qsec         vs          am       gear        carb
## mpg   0.41868403  0.6640389  0.59983243  0.4802848 -0.55092507
## cyl  -0.59124207 -0.8108118 -0.52260705 -0.4926866  0.52698829
## disp -0.43369788 -0.7104159 -0.59122704 -0.5555692  0.39497686
## hp   -0.70822339 -0.7230967 -0.24320426 -0.1257043  0.74981247
## drat  0.09120476  0.4402785  0.71271113  0.6996101 -0.09078980
## wt   -0.17471588 -0.5549157 -0.69249526 -0.5832870  0.42760594
## qsec  1.00000000  0.7445354 -0.22986086 -0.2126822 -0.65624923
## vs    0.74453544  1.0000000  0.16834512  0.2060233 -0.56960714
## am   -0.22986086  0.1683451  1.00000000  0.7940588  0.05753435
## gear -0.21268223  0.2060233  0.79405876  1.0000000  0.27407284
## carb -0.65624923 -0.5696071  0.05753435  0.2740728  1.00000000
```

This correlation matrix allows us to identify strong correlations between the following variables : cyl & disp, cyl & hp, cyl & vs, disp & wt. With the support of the plot of appendix 1 (pair plots including those 4 variables), decision is taken to suppress the cyl & disp variables as they are highly correlated with hp, ws and vs. This makes sense because bigger cars will have bigger engines with more cylinders… The other variables will be kept because they can contribute to explain variability in the model.
In addition, moderate correlation values between our response variable and qsec, gear and carb, we will not consider those variables as well.

Model definition

In this step, we will fit a set of models by conducting a forward approach (step by step nested approach with new variables added at each step). mpg will be our response variable and as our questions are related to the impact of transmission mode, the am variable will be embedded in each model. Our categorical variables will be handled as factors. Anova will be applied to select the best model (nested likelihood ratio tests).

```
fit0 <- lm(mpg~factor(am),my_mtcars)
fit1 <- lm(mpg~factor(am)+hp,my_mtcars)
fit2 <- lm(mpg~factor(am)+drat,my_mtcars)
fit3 <- lm(mpg~factor(am)+wt,my_mtcars)
fit4 <- lm(mpg~factor(am)+factor(vs),my_mtcars)
fit5 <- lm(mpg~factor(am)+hp+factor(vs),my_mtcars)
fit6 <- lm(mpg~factor(am)+hp+factor(vs)+drat+wt,my_mtcars)

anova(fit0,fit5)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ factor(am) + hp + factor(vs)
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1     30 720.90
## 2     28 218.88  2    502.02 32.11 5.658e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As a second step, we control the evolution of the coefficient associated to the “am” variable.

```
rbind(summary(fit0)$coef[2,], summary(fit1)$coef[2,], summary(fit2)$coef[2,], summary(fit3)$coef[2,], summary(fit
4)$coef[2,], summary(fit5)$coef[2,],summary(fit6)$coef[2,])
```

```
##         Estimate Std. Error     t value     Pr(>|t|)
## [1,]  7.24493927   1.764422  4.10612698 2.850207e-04
## [2,]  5.27708531   1.079541  4.88826953 3.460318e-05
## [3,]  2.80706095   2.282159  1.23000231 2.285814e-01
## [4,] -0.02361522   1.545645 -0.01527855 9.879146e-01
## [5,]  6.06666667   1.274842  4.75875870 4.958115e-05
## [6,]  5.29853680   1.037569  5.10668299 2.071740e-05
## [7,]  2.08595742   1.604920  1.29972713 2.051005e-01
```

Finally we select fit5 as our best model as this model is the one which minimizes the std. Error for the coefficient associated to the am variable. Diagnosis elements associated to the model are given in appendix number 3 and allow us to conclude that the model fit appears to be valid.

Interpretations

On average, with other variables being fixed, compared to an automatic transmission, manual transmission increases mpg by 5.29. This value is statistically significant as the Pr associated to this coefficient is < to 0.05.

The confidence interval associated to the coefficient (manual transmission) is estimated below :

```
confint(fit5)[2,]
```

```
##    2.5 %   97.5 %
## 3.173173 7.423901
```

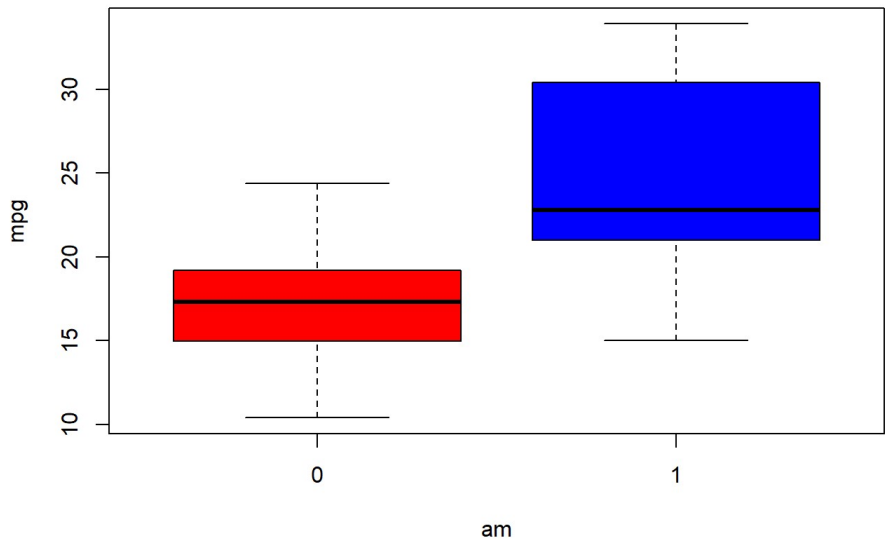
As this confidence interval doesn't include zero, We can conclude that manual transmission mode is more efficient than automatic transmission !

Appendixes - Plots

Appendix 1 - EDA mpg ~ auto / manu transmission

The main concern is related to identification of links between transmission mode (auto / manu) and mpg. This plot is used to visualize mpg distributions vs transmission categories.

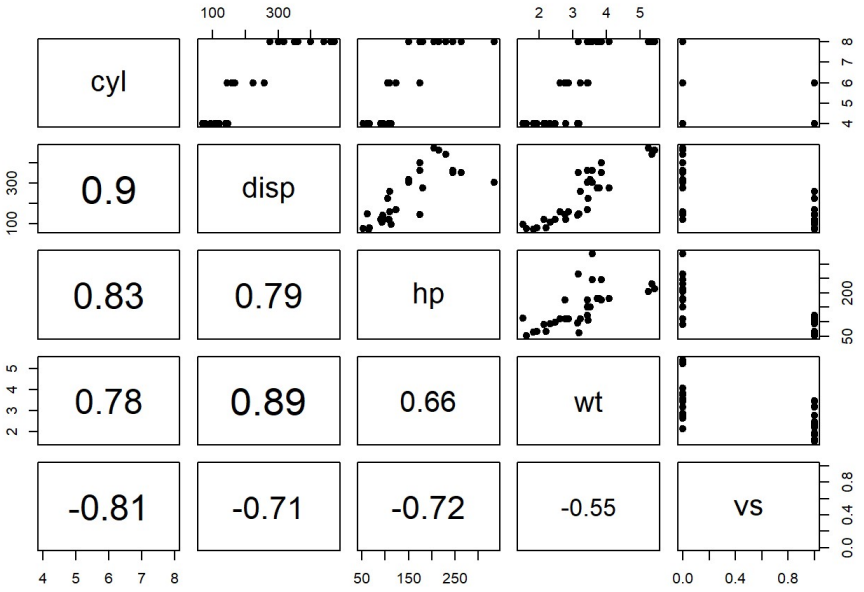
```
#ggplot boxplot not printed correctly with Knitr, using basing plotting instead.
boxplot(mpg ~ am, data = my_mtcars, col=(c("red","blue")))
```



Appendix 2 - pair plots

```
# Function used to dispkay correlation coefficients on pair plots
panel.cor <- function(x, y, ...)
{
par(usr = c(0, 1, 0, 1))
txt <- as.character(format(cor(x, y), digits=2))
text(0.5, 0.5, txt, cex = 3* abs(cor(x, y)))
}

my_mtcars %>% select (cyl,disp,hp,wt,vs) %>% pairs(lower.panel=panel.cor,pch = 19)
```

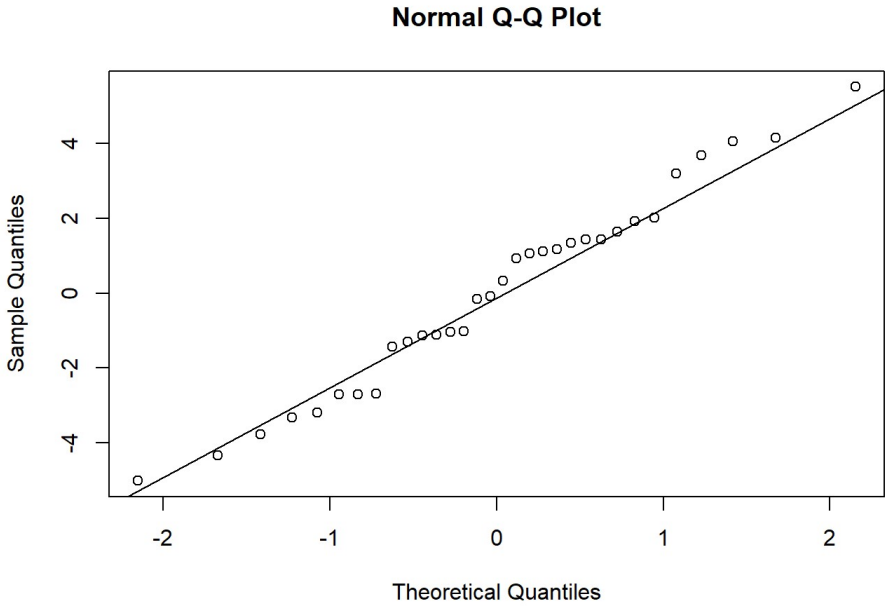


Appendix 3 - Model Diagnosis

As we selected model fit5 as our best model, let's perform some diagnosis plots.

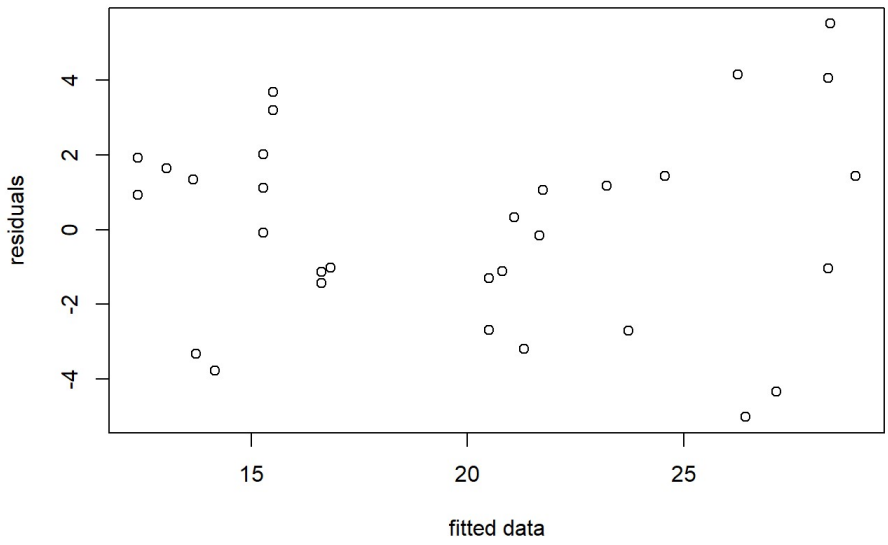
Normality condition for the residuals

```
qqnorm(fit5$residuals)
qqline(fit5$residuals)
```



The normality condition is satisfied.

```
# Valeurs absolues
plot((fit5$residuals) ~ fit5$fitted, xlab ="fitted data", ylab="residuals")
```



The constant variability of the residuals versus the fitted data condition is satisfied.