ProjectRegressionModels

Alejandro Coy

2019-02-23

## Peer-graded Assignment: Regression Models Course Project

### Executive summary

The main objective of this projet is to apply the regressions model to answer the following questions :

1. “Is an automatic or manual transmission better for MPG”?
2. “Quantify the MPG difference between automatic and manual transmissions” ?

Using exploratory data anlysis the difference between automatic and manual transmission cars was evident. Furthermore a t test was perfomed for the two groups showing a significant difference between the groups, showing the automatic cats achived less MPG than manual cars.

For quantifiy the difference a linear regrtession model was used. Thre models were tested with differente explanatory variables. With the best model it was determined that to have a manual transmission increase 1.63 MPG when all the other variable are held constant.

### Exploratory Data Analysis

The dataset is preload in Rstudio. The columns with categorical value were changd to factor using the mutate function.

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)   
library(GGally)

##   
## Attaching package: 'GGally'

## The following object is masked from 'package:dplyr':  
##   
## nasa

data <- mtcars %>%  
 mutate\_at(c("cyl","gear","carb","am","vs"), as.factor)  
levels(data$am) <- list(automatic="0", manual="1")

A serie of plots using ggpairs and gplot were done to see the relationship between MPG and the varaible (see APEENDIX)

The first exploration indicates that manual cars yield higher MPG.

### T-Test

In order to perform a t-test the data was divided depending on the type of transmission. Then the t.test function was performed:

t.test(automatic$mpg,manual$mpg,paired = FALSE)

##   
## Welch Two Sample t-test  
##   
## data: automatic$mpg and manual$mpg  
## t = -3.7671, df = 18.332, p-value = 0.001374  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -11.280194 -3.209684  
## sample estimates:  
## mean of x mean of y   
## 17.14737 24.39231

As can be seen there is a significant diffrente with a lower mean for automatic cars. The interval confidence does not contain 0 and p value is lowwer than 0.05

### Regression models

In order to quantify the difference between automatic and manual transmissions three linear models were used.

#### Model 1

Model 1 just consider the type of tranmission (am) as explanatory varaible

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 17.147368 1.124603 15.247492 1.133983e-15  
## ammanual 7.244939 1.764422 4.106127 2.850207e-04

The coefficient of the am variable indicates that the use of manual transmission increase 7.245 MPG, However, the R-squared is low indicating the change in transmission can just explain 35.9% the variation of the MPG.

#### Model2

The model 2 take in consideration the variable that seem to have influence according to the pair plot from the EDA:

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 30.02707736 6.185177080 4.854684 4.500893e-05  
## hp -0.03637325 0.009814491 -3.706076 9.579590e-04  
## drat 0.98101801 1.377101246 0.712379 4.823413e-01  
## ammanual 1.57852084 1.559281497 1.012339 3.203628e-01  
## wt -2.72609168 0.937791027 -2.906929 7.209042e-03

The R-square for this model is 0.8428 with Adjusted R-squared : 0.8196. This indicates a better model. The residual plot are distributed around 0 which indicates a good fitting model.

### Model3

Finlaly model 3 take in considreation all the variables

model3 <- lm(mpg ~., data= data)

The R-square for this model is 0.8931, but an Adjusted R-squared : 0.779. This is an indication of overfitting of the model. Finally we can compare between the models using the anova anlysis:

anova(model1,model2,model3)

## Analysis of Variance Table  
##   
## Model 1: mpg ~ am  
## Model 2: mpg ~ hp + drat + am + wt  
## Model 3: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 30 720.90   
## 2 27 176.96 3 543.93 22.5880 8.124e-06 \*\*\*  
## 3 15 120.40 12 56.56 0.5872 0.821   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

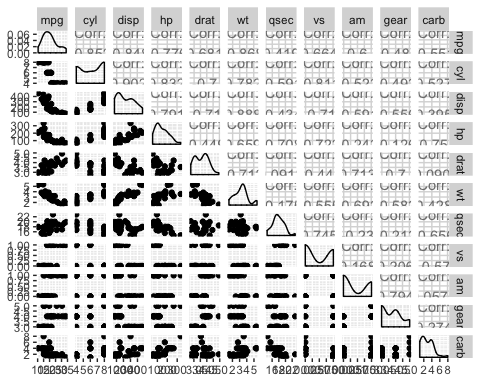
From this results we can conclude there is not need to add all the variables in model 3 since there is not significant difference in the predections. However in comparison with model 1, model 2 does a significant better job (p-value lower than 0.05).

### 

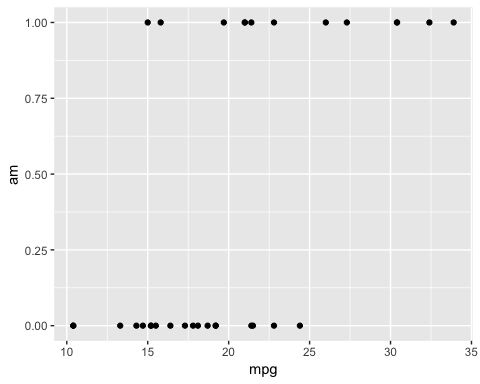
### APPENDIX

GRID PLOTS

ggpairs(mtcars)

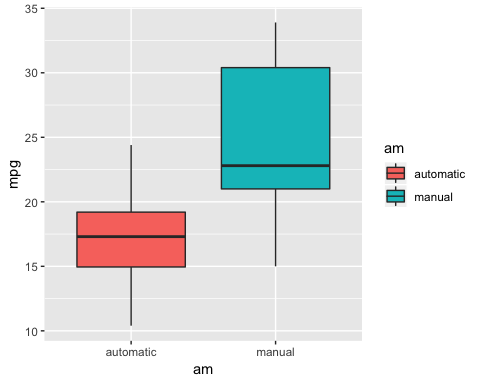


ggplot(data = mtcars, aes(y =am, x = mpg))+  
 geom\_point()



MEAN MPG FOR DIFFERENT TYPE OF TRANSMISSION

ggplot(data = data)+  
 aes(x = am, y = mpg,fill=am)+  
 geom\_boxplot()



REISIDUAL PLOT FOR MODEL 2

qplot(predict(model2), resid(model2))

