Peer-graded Assignment: Regression Models Course Project

Santiago ruiz navas

11/26/2019

Contents

Executive summary

This report presents the answers to two research questions:
* Is an automatic or manual transmission better for Milles (US) per Gallon (MPG)
* Quantify the MPG difference between automatic and manual transmissions
These questions will be answered using the dataset mtcars and a three step process:
* Exploratory analysis: review the characteristics of the data and obtan the corelations between the predictor AM and the other predictors
* Nested models: Nest various models and evalaute them using anova and saphiro tests
* Inference: Calcualte for the 95% interval of the coeficient corresponding to AM in the selected model
In conclusion:
* The use of automatic or manual transmision does not have a significant effect in MPG.
* The avearage different quantity of mpg related to using automatic or manual tranmision was xxx with 95% interval of xx.
* The consideration of other variables available in the dataset in the models seriously affects the coeficient of the automatic and manual tranmision in the model.

Exploratory analisys

Loading data & libraries

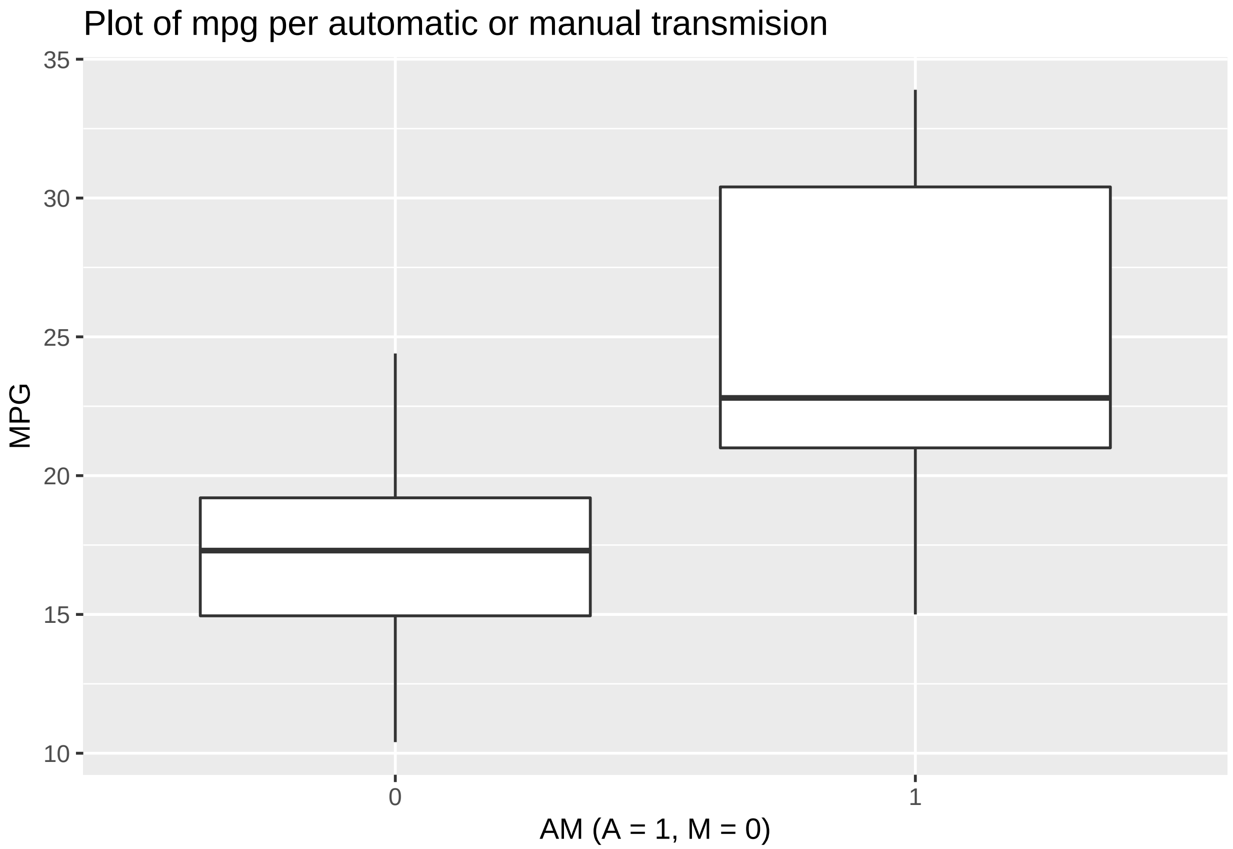
```
library(ggplot2)
data("mtcars")
```

Summary of data

```
summary(mtcars)
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
##        am              gear            carb      
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
##  Median :0.0000   Median :4.000   Median :2.000  
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812  
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000  
```

In the boxplot bellow the values for mpg given a automatic or manual transmision are presented:



From the graph above it is possible to see a difference between using or not auotmatic transmision and the value of MPG. However, is this difference significant regaring the other variables in the dataset?

Quick one variable analysis

Preparing the predictor variable as a factor

```
mtcars$am <- as.factor(mtcars$am)
mdl.project <- lm(mpg ~ am, mtcars)
```

Looking at the slope to see the difference in mpgs between using auto or manual transmision

```
summary(mdl.project)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am1          7.244939   1.764422  4.106127 2.850207e-04
```

The coeficient seem correctand the difference significative with manual mpg goes up. Now the model is tested using the anova and shapiro tests to see if this conclusion is based in a solid model.

```
anova(mdl.project)# okay
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value   Pr(>F)
## am         1 405.15  405.15   16.86 0.000285 ***
## Residuals 30 720.90   24.03
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# checking for the normality of the coeficients
shapiro.test(mdl.project$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mdl.project$residuals
## W = 0.98208, p-value = 0.8573
```

```
# can not deny the null hipothesis then the residuals are normal! win
```

Finally lets see the confidence interval for the calculated coeficient

```
coef.statistics <- summary(mdl.project)$coef
B1.interval <- coef.statistics[2,1] +c(-1,1)*qt(.975, df = mdl.project$df)*coef.statistics[2,2]#
```

The 95% interval for B1 was above cero. Therfore, is signficative.

Nested model analysis

The strategy for model selection is nested analysis. I will check the corelations between the predictor AM and the other variables of the dataset, and in the order of the magnitud of their correlations will create nested models and evaluate them.The justification for this corelation-based-policy is that, the more correlation another predictor has with AM the more impact it will have on its regressed coeficient. Furthremore, the criteria to select a model is based on how "strong" (p-value) they pass the anova and shapiro tests. * The evaluation of a model consits in running anova and shapiro tests. * The "strengh" of a model is proportional to the significance in which it passes the tests. e.g., given mdl1 and mdl2, if mdl 1 passes anova and shapiro with a 95% of significance and mdl2 with 90% then mdl1 is stronger than mdl2.

getting the correlations of AM

```r
data("mtcars")
mtcars.cor <- cor(mtcars)
am.cor <- sort(abs(mtcars.cor[,"am"]), decreasing = T)
```

The order showed by the vector am is followed to test the nested models

Testing the nested models

Three iterations were made. However, only the first will be shown because of space constrain

Results of nested models

After the three iterations the model mpg ~ am + gear + drat + wt + cyl was selected with anova p-value of xxx and saphiro p-value of xx. Bellow the results of the three iterations is explained * In the first iteration all variables were nested, after the evalaution the varible disp was removed. * In the second iteration hp was removed. * In the third iteration qsec was removed and the best model based on its anova and saphiro tests results was selected

Inference

Calcualting the inference interval for the B1 B1 of the selected model was 1.301 with interval of -2.46 and 5.06 (mpgs) The interval includes cero, therefore it is considered that the difference is not significant.

Conclusions

- ANSWER TO QUESTION 1: The interval of the B1 for the selected model was tested at a 95% t-test and it contained cero. Therefore, taking into account other variables in the dataset, AM does not have a significant effect on mpg
- ANSWER TO QUESTION 2: B1 1.301mpg increse when using manual over auto, with interval of -2.46 and 5.06 (mpgs)

Limitations

In this approach only three iterations of one set of possible combinations of the variables was done, in a future study all the other possible combinations of the variables can be explored to find the best possbile model and with re-evaluate the answer to these question.

Assumptions

The assumptions to compare binary values using linear regressions are that * The samples of the compared values were idepedent iid * The variances of the compared variables were equal