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 obtain the correlations between the predictor AM and the other predictors
* Nested models: Nest various models and evaluate them using ANOVA and Shapiro tests
* Inference: Calculate for the 95% interval of the coefficient corresponding to am in the selected model
In conclusion:
* The use of automatic or manual transmission **does not have a significant effect on mpg**.
* The average difference quantity of mpg related to using manual over automatic transmission was **1.301 with a 95% interval of -2.46 and 5.06**.
* The present analysis has the limitation that it assumes constant variance of the values of am to mpg and that the data follow normal distributions. The two previous assumptions might not be real in the case of the mtcars dataset. Therefore, the conclusions from the inference analysis might be wrong.

Exploratory analisys

Heads of data

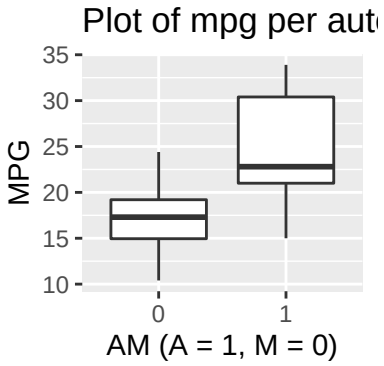
```
library(ggplot2)
data(mtcars)
head(mtcars)
```

```
##                   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
```

In the boxplot below, it is possible to see the values of mpg for automatic or manual transmission:

```
p <- ggplot(mtcars, aes(x=as.factor(mtcars$am), y=mtcars$mpg)) +
  geom_boxplot()+ labs(title="Plot of mpg per automatic or manual transmision", x="AM (A = 1, M = 0)",
                       y = "MPG")
p # in mpgs when using automatic vs manual
```



From the graph above, it is possible to see a difference between using auotmatic or manual transmission and the value of MPG. However, one question remains, is this difference significant regarding the other variables in the dataset?

Nested model analysis

The strategy for model selection is nested analysis. I will check the correlations between the predictor AM and the other variables of the dataset, and in the order of the magnitude of their correlations, will create nested models and evaluate them. The justification for this correlation-based-policy is that the more correlated the other predictors are to AM, the more impact they will have on its regressed coefficient. Furthermore, 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 consists of running ANOVA and Shapiro tests.
* The "strength" of a model is proportional to the significance in which it passes the tests. e.g., given mdl1 and mdl2, if mdl 1 passes the ANOVA and Shapiro tests with 95% significance and mdl2 with 90%, then mdl1 is stronger than mdl2.

Getting the correlations of am

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

```
##         am       gear       drat         wt        mpg       disp
## 1.00000000 0.79405876 0.71271113 0.69249526 0.59983243 0.59122704
##        cyl         hp       qsec         vs       carb
## 0.52260705 0.24320426 0.22986086 0.16834512 0.05753435
```

The nested models are tested in the order showed by the vector am

Testing the nested models

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

Create the models

```
mdl.project  <- lm(mpg ~ am, mtcars)
mdl.project1 <- lm(mpg ~ am+gear, mtcars)# gear
mdl.project2 <- lm(mpg ~ am+gear+drat, mtcars)# drat
mdl.project3 <- lm(mpg ~ am+gear+drat+wt, mtcars) #wt
mdl.project4 <- lm(mpg ~ am+gear+drat+wt+disp, mtcars) #disp
mdl.project5 <- lm(mpg ~ am+gear+drat+wt+disp+cyl, mtcars) #cyl
mdl.project6 <- lm(mpg ~ am+gear+drat+wt+disp+cyl+hp, mtcars) #hp
mdl.project7 <- lm(mpg ~ am+gear+drat+wt+disp+cyl+hp+qsec, mtcars) #qsec
mdl.project8 <- lm(mpg ~ am+gear+drat+wt+disp+cyl+hp+qsec+carb, mtcars) #carb
mdl.project9 <- lm(mpg ~ am+gear+drat+wt+disp+cyl+hp+qsec+carb+vs, mtcars) #vs
```

Evaluating saphiro test

```
s <- c(shapiro.test(mdl.project$residuals)$p.value,
  shapiro.test(mdl.project1$residuals)$p.value,  shapiro.test(mdl.project2$residuals)$p.value,
  shapiro.test(mdl.project3$residuals)$p.value,  shapiro.test(mdl.project4$residuals)$p.value,
  shapiro.test(mdl.project5$residuals)$p.value,  shapiro.test(mdl.project6$residuals)$p.value,
  shapiro.test(mdl.project7$residuals)$p.value,  shapiro.test(mdl.project8$residuals)$p.value,
  shapiro.test(mdl.project9$residuals)$p.value)
names(s) <- c("m","m1","m2","m3","m4","m5","m6","m7","m8","m9")
s # m4 is bad
```

```
##          m         m1         m2         m3         m4         m5
## 0.85734421 0.87968915 0.94129300 0.33171114 0.01567496 0.32055995
##         m6         m7         m8         m9
## 0.07968890 0.10205830 0.22605866 0.22614893
```

The p-values of the Shapiro test show that for m4 the null hypothesis does not hold. Therefore, m4 is not adequate.
Evaluating anova test

```
a <- anova(mdl.project, mdl.project1,  mdl.project2, mdl.project3, mdl.project4,
           mdl.project5, mdl.project6, mdl.project7, mdl.project8, mdl.project9)
a$`Pr(>F)`[5]# m4 is bad
```

```
## [1] 0.06416311
```

Again m4 does not pass the test. Therefore, the variable added in model 4 (disp) is removed from the analysis.

Results of nested models

After the three iterations the model **mpg ~ am + gear + drat + wt + cyl** was selected with ANOVA **p-value of 0.00209** and **Saphiro p-value of 0.2862**. Bellow the results of the three iterations are explained
* In the first iteration, all variables were nested. After the evaluation, **the variable disp was removed**.
* In the second iteration **hp was removed**.
* In the third iteration the model **mpg ~ am + gear + drat + wt + cyl** was selected based on its ANOVA and Shapiro tests results

Inference

Calculating the inference interval for the estimated coefficient for am:
The coefficient for am in the selected model was **1.301 with an interval of -2.46 and 5.06 (mpgs)**
The interval includes cero. Therefore, the difference between using manual over automatic transmission on mpgs is not significant.

Conclusions

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

Limitations

In this report, I did three iterations of model evaluation; in a future study, other combinations can be explored to find the best possible model, and with it, re-evaluate the answers to these questions.

Assumptions

The assumptions to compare binary values using linear regressions are
* The values for the variables of mpg and am followed a normal distributions
* The variances of the set am concerning mpg were equal

ANNEX

Iteration 2

```
mdl.project   <- lm(mpg ~ am, mtcars)
mdl.project1  <- lm(mpg ~ am+gear, mtcars)# gear
mdl.project2  <- lm(mpg ~ am+gear+drat, mtcars)# drat
mdl.project3  <- lm(mpg ~ am+gear+drat+wt, mtcars) #wt
mdl.project10 <- lm(mpg ~ am+gear+drat+wt+cyl, mtcars) #cyl
mdl.project11 <- lm(mpg ~ am+gear+drat+wt+cyl+hp, mtcars) #hp
mdl.project12 <- lm(mpg ~ am+gear+drat+wt+cyl+hp+qsec, mtcars) #qsec
mdl.project13 <- lm(mpg ~ am+gear+drat+wt+cyl+hp+qsec+carb, mtcars) #carb
mdl.project14 <- lm(mpg ~ am+gear+drat+wt+cyl+hp+qsec+vs, mtcars) #vs
```

```
s <- c(shapiro.test(mdl.project$residuals)$p.value,
       shapiro.test(mdl.project1$residuals)$p.value,shapiro.test(mdl.project2$residuals)$p.value,
       shapiro.test(mdl.project3$residuals)$p.value,shapiro.test(mdl.project10$residuals)$p.value,
       shapiro.test(mdl.project11$residuals)$p.value,shapiro.test(mdl.project12$residuals)$p.value,
       shapiro.test(mdl.project13$residuals)$p.value, shapiro.test(mdl.project14$residuals)$p.value)
names(s) <- c("m","m1","m2","m3","m10","m11","m12","m13","m14")

a <- anova(mdl.project, mdl.project1,
           mdl.project2, mdl.project3,
           mdl.project10,mdl.project11,# do not pass the anova hp
           mdl.project12,mdl.project13,
           mdl.project14           )
# Another iteration removing disp and hp
```

Iteration 3

```
mdl.project   <- lm(mpg ~ am, mtcars)
mdl.project1  <- lm(mpg ~ am+gear, mtcars)# gear
mdl.project2  <- lm(mpg ~ am+gear+drat, mtcars)# drat
mdl.project3  <- lm(mpg ~ am+gear+drat+wt, mtcars) #wt
mdl.project10 <- lm(mpg ~ am+gear+drat+wt+cyl, mtcars) #cyl
mdl.project15 <- lm(mpg ~ am+gear+drat+wt+cyl+qsec, mtcars) #qsec
mdl.project16 <- lm(mpg ~ am+gear+drat+wt+cyl+qsec+carb, mtcars) #carb
mdl.project17 <- lm(mpg ~ am+gear+drat+wt+cyl+qsec+vs, mtcars) #vs

s <- c(shapiro.test(mdl.project$residuals)$p.value,
       shapiro.test(mdl.project1$residuals)$p.value,shapiro.test(mdl.project2$residuals)$p.value,
       shapiro.test(mdl.project3$residuals)$p.value,shapiro.test(mdl.project10$residuals)$p.value,
       shapiro.test(mdl.project15$residuals)$p.value,shapiro.test(mdl.project16$residuals)$p.value,
       shapiro.test(mdl.project17$residuals)$p.value)
names(s) <- c("m","m1","m2","m3","m10","m15","m16","m17")

a <- anova(mdl.project,mdl.project1, mdl.project2, mdl.project3,
           mdl.project10, mdl.project15,# do not pass the anova hp
           mdl.project16, mdl.project17
)
```

Confidence interval for the estimated coeficient in the selected model

```
mdl.project10.coef.statistics <- summary(mdl.project10)$coef
mdl.project10.coef.statistics.B1.int <-
  mdl.project10.coef.statistics[2,1] +c(-1,1)*qt(.975, df = mdl.project10$df)*mdl.project10.coef.statis
mdl.project10.coef.statistics.B1.int
```