Regression Model Project

By Angie Marchany-Rivera

08/03/2021

Executive Summary

In this analysis we are interested in finding which type of transmission (coded as 'am' where 0 = automatic, 1 = manual) is more efficient in terms of miles per gallon (mpg) using the mtcars dataset in R.The exploratory analysis of the data suggested that manual transmissions yield higher average mpg than automatic transmissions. This relationship was further analyzed using nested multivariable linear regression models. The variables cyl and disp were removed from the linear regression due to high collinearity with other predictors. Based on this regression it was concluded that manual transmissions do not yield a statistically significant difference in mpg compared to automatic transmissions.

Loading the mtcars data and performing some basic exploratory data analysis

```
library(usdm)
data(mtcars); str(mtcars)
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

The mtcars dataset contains 32 observation and 11 variables. This data was further summarized based on the transmission type to explore its relationship with the other variables. It is worth noting that manual transmission yielded higher miles per gallons than automatic transmission (see Appendix A - Table 1: mpg mean values). Linear regression models will be created to analyze the relationship between transmission type and mpg. This analysis will assume that the variable mpg is the desired outcome and the rest of the variables are predictors.

It can be expected that some of the predictors have high correlation values (see Appendix A - Plot 1). Variable correlation was further analyzed using variance inflation factors.

```
vifcor(subset(mtcars,select = (-mpg)),th = 0.8) #correlation treshold set to 0.8
...
## 2 variables from the 10 input variables have collinearity problem:
## disp cyl
## After excluding the collinear variables, the linear correlation coefficients ranges between:
## min correlation ( carb ~ am ):  0.05753435
## max correlation ( gear ~ am ):  0.7940588
...
```

The vifcor function excludes any predictor with a VIF > 10 since the information that these predictors provide about the response is redundant in the presence of the other variables. Using the vif values as a

guide, the variables disp and cyl were removed from the linear regression model. The effect of removing these predictors were analyzed using anova and the Shaphiro-Wilk test. The results from these tests are shown on the next session.

Multivariable Regression Models

```
# Creating 3 nested linear models:
fit1 <- lm(mpg~factor(am),mtcars)
fit2 <- lm(mpg~factor(am)+. -am,mtcars)
fit3 <- lm(mpg~factor(am) +. -disp -cyl -am, mtcars)
# Comparing the 3 models unsing anova
anova(fit1,fit3,fit2)
...
## Analysis of Variance Table
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     23 151.48  7    569.42 11.5819 5.772e-06 ***
## 3     21 147.49  2      3.98  0.2834    0.7561
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
...
```

The p-values of the anova test shows that the most significant model is fit3, p-value = 5.772-06. Checking that the residuals of fit3 follow normality:

```
shapiro.test(fit3$residuals)
...
##  Shapiro-Wilk normality test
## W = 0.96495, p-value = 0.3727
...
```

The Shapiro-Wilk p-value of 0.3727 fails to reject normality of the residuals. Normality of the residuals can also be verified by the linearity of the residuals Q-Q plot (see Appendix B - Plots). We can also compare the adjusted R-squared values of the 3 models to confirm that fit3 is the best fit (see Appendix B - Summary for the complete output of the function summary(fit3)).

```
cbind(summary(fit1)$adj.r.squared,summary(fit2)$adj.r.squared,summary(fit3)$adj.r.squared )
##           [,1]      [,2]      [,3]
## [1,] 0.3384589 0.8066423 0.8186912
```

Conclusion

Based on the analysis performed, the coefficients of the linear model fit3 can be used to summarize the relationship between transmission type and mpg. The coefficient for manual transmission estimates that manual transmissions yield, in average, 2.4 mpg more than automatic transmissions. However, the p-value of this estimate is greater than 0.05 which makes it not statistically significant.The rest of the coefficients indicate that weight is the only predictor that has a statistically significant impact on mpg when cylinder and displacement are held constant.

Appendix A - Exploratory Graphs and Table

Plot 1: Relationship between the variables:

```
plot(mtcars, pch = 16, col = "blue")
```

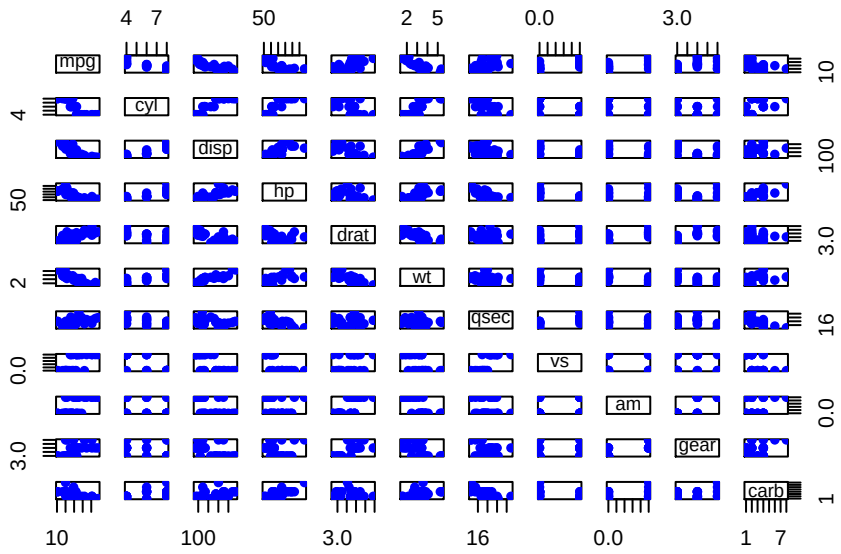

Table 1: mtcars dataset - Variable descriptions and summary by transmission type

```
library(dplyr); library(gtsummary); library(kableExtra)
mtcars %>% tbl_summary(by=am, label=list(mpg ~ "MPG - Miles/(US) gallon",
                                         cyl ~ "cyl - Number of cylinders ",
                                         disp ~ "disp - Displacement (cu.in.)",
                                         hp ~ "hp - Gross horsepower",
                                         drat ~ "drat - Rear axle ratio",
                                         wt ~ "wt - Weight (1000 lbs)",
                                         qsec ~ "qsec - 1/4 mile time",
                                         vs ~ "vs - Engine (0 = V-shaped, 1 = straight)",
                                         gear ~ "gear - Number of forward gears",
                                         carb ~ "carb - Number of carburetors"),
       type=all_continuous() ~ "continuous2",statistic = all_continuous() ~"{mean} ({sd})")%>%
       as_kable_extra() %>% row_spec( c(1,3,7,9,11,13,15,17,18,22),hline_after=TRUE,
                                      bold=TRUE, color = "white",background = "black")%>%
       add_header_above(c(" "=1, "Automatic" = 1, "Manual" = 1),bold=TRUE)%>%
       kable_styling(bootstrap_options = "striped", latex_options = "hold_position",
                     full_width = F, font_size =12) %>% save_kable("./table1.png",zoom =5)
```

	Automatic	Manual
Characteristic	**0, N = 19**	**1, N = 13**
MPG - Miles/(US) gallon		
Mean (SD)	17.1 (3.8)	24.4 (6.2)
cyl - Number of cylinders		
4	3 (16%)	8 (62%)
6	4 (21%)	3 (23%)
8	12 (63%)	2 (15%)
disp - Displacement (cu.in.)		
Mean (SD)	290 (110)	144 (87)
hp - Gross horsepower		
Mean (SD)	160 (54)	127 (84)
drat - Rear axle ratio		
Mean (SD)	3.29 (0.39)	4.05 (0.36)
wt - Weight (1000 lbs)		
Mean (SD)	3.77 (0.78)	2.41 (0.62)
qsec - 1/4 mile time		
Mean (SD)	18.18 (1.75)	17.36 (1.79)
vs - Engine (0 = V-shaped, 1 = straight)	**7 (37%)**	**7 (54%)**
gear - Number of forward gears		
3	15 (79%)	0 (0%)
4	4 (21%)	8 (62%)
5	0 (0%)	5 (38%)
carb - Number of carburetors		
1	3 (16%)	4 (31%)
2	6 (32%)	4 (31%)
3	3 (16%)	0 (0%)
4	7 (37%)	3 (23%)
6	0 (0%)	1 (7.7%)
8	0 (0%)	1 (7.7%)

Appendix B - Linear Model summary and plot

Summary of fit3:

```
##
## Call:
## lm(formula = mpg ~ factor(am) + . - disp - cyl - am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8187 -1.3903 -0.3045  1.2269  4.5183
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.80810   12.88582   1.072   0.2950
## factor(am)1  2.42418    1.91227   1.268   0.2176
## hp          -0.01225    0.01649  -0.743   0.4650
## drat         0.88894    1.52061   0.585   0.5645
## wt          -2.60968    1.15878  -2.252   0.0342 *
## qsec         0.63983    0.62752   1.020   0.3185
## vs           0.08786    1.88992   0.046   0.9633
## gear         0.69390    1.35294   0.513   0.6129
## carb        -0.61286    0.59109  -1.037   0.3106
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.566 on 23 degrees of freedom
## Multiple R-squared:  0.8655, Adjusted R-squared:  0.8187
## F-statistic:  18.5 on 8 and 23 DF,  p-value: 2.627e-08
```

Plots of the residuals for fit3: