Motor Trend, Regression Models Course Project

Irina White

29/05/2021

Introduction

This model has been constructed to answer the two key questions using regression models and exploratory data analyses: *Is an automatic or manual transmission better for MPG? Quantify the MPG difference between automatic and manual transmissions?*

Summary

For the purpose of this analysis the data set **mtcars** has been used with the variable MPG as a desired outcome and the key am variable as a regressor,with additional regressors added to the multivariable regression analysis.

The findings suggest that manual transmission tends to result in more mileage (as a result of the single variable t-test), further analysis based on quantifying the difference between transmission with additional regressors, leads to the conclusion that other factors affecting mileage along with the type of the trasmission are weight(wt), gross horsepowers(hp), and number of cylinders (cyl).

```
head(mtcars, 8)
```

```
##                   mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
## Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
```

As it can be observed from the quick overview of the data, the following variables can be used as factors (categorical data): cyl(4,6,8), gear(3,4,5), carb(1,2,3,4,6,8), vs (0,1) and am (0,1).

```
#convert continuous variables into categoricals in data set mtcarsf
mtcarsf<-mtcars
cols <- c("cyl", "vs", "am", "gear",'carb')
mtcarsf[cols] <- lapply(mtcarsf[cols], factor)
```

Is an automatic or manual transmission better for MPG?

The simple T-test is to be used to asses the influence of the transmission on mileage. The variable that represents transmission is **am** (0 = automatic, 1 = manual)

Without taking into account other variables the T-test fails to support the null hypothesis that there is no significant difference between Automatic or Manual transmission effect on mpg.

```
t.test(mtcars$mpg~mtcars$am)
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars$mpg by mtcars$am
## 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 in group 0 mean in group 1
##        17.14737        24.39231
```

The outcome suggests that **manual transmission results in higher mileage**. (p =0.0014, automatic mean = 17.1 miles, manual mean = 24.4 miles). The boxplot can be found in Appendix section.

Quantify the MPG difference between automatic and manual transmissions?

1. Correlation

Brief correlation analysis shows that mpg has higher level of correlation with wt, cyl, disp, and hp. However wt and disp is highly correlated between each other therefore one should be omitted to avoid

```
round(cor(mtcars, use='everything', method='pearson'),2)
```

```
##        mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
## mpg   1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
## cyl  -0.85  1.00  0.90  0.83 -0.70  0.78 -0.59 -0.81 -0.52 -0.49  0.53
## disp -0.85  0.90  1.00  0.79 -0.71  0.89 -0.43 -0.71 -0.59 -0.56  0.39
## hp   -0.78  0.83  0.79  1.00 -0.45  0.66 -0.71 -0.72 -0.24 -0.13  0.75
## drat  0.68 -0.70 -0.71 -0.45  1.00 -0.71  0.09  0.44  0.71  0.70 -0.09
## wt   -0.87  0.78  0.89  0.66 -0.71  1.00 -0.17 -0.55 -0.69 -0.58  0.43
## qsec  0.42 -0.59 -0.43 -0.71  0.09 -0.17  1.00  0.74 -0.23 -0.21 -0.66
## vs    0.66 -0.81 -0.71 -0.72  0.44 -0.55  0.74  1.00  0.17  0.21 -0.57
## am    0.60 -0.52 -0.59 -0.24  0.71 -0.69 -0.23  0.17  1.00  0.79  0.06
## gear  0.48 -0.49 -0.56 -0.13  0.70 -0.58 -0.21  0.21  0.79  1.00  0.27
## carb -0.55  0.53  0.39  0.75 -0.09  0.43 -0.66 -0.57  0.06  0.27  1.00
```

2. Variance Inflation Factor (library 'car')

```
library(car)
```

```
## Loading required package: carData
```

```
fit1<-lm(mpg~am+wt+hp+cyl+disp, mtcars)
fit2<-lm(mpg~am+wt+hp+cyl, mtcars)
fit3<-lm(mpg~am+wt+hp, mtcars)
fit4<-lm(mpg~am+wt, mtcars)
fit5<-lm(mpg~am, mtcars)
round(rbind(vif(fit1), c(vif(fit2),0), c(vif(fit3),0,0), c(vif(fit4),0,0,0)),3)
```

```
##         am    wt    hp   cyl   disp
## [1,] 2.553 6.079 4.502 7.209 10.401
## [2,] 2.546 3.988 4.310 5.334  0.000
## [3,] 2.271 3.775 2.088 0.000  0.000
## [4,] 1.921 1.921 0.000 0.000  0.000
```

VIF analysis shows that disposition has a very high correlation rate, as soon as the variable is omitted from the analysis, the relative importance of correlated variables is excluded as well and the VIF rates are moderate.

3. Linear Regression Model Simple vs Multivariable

Therefore it is worth to consider two models: - first the one variable model that includes only am - multivariable model that includes: am, wt, hp and cyl

```
summary(fit5)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am             7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

The cars with the automatic transmission has on average 17.15mpg, and the cars with the manual transmission result in 17.15+7.24= 24.49mpg, the p-value <0.05, the adj.r.squared is 33.9% thus covers almost 34% of variance.

```
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + hp + cyl, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4765 -1.8471 -0.5544  1.2758  5.6608
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.14654    3.10478  11.642 4.94e-12 ***
## am           1.47805    1.44115   1.026   0.3142
## wt          -2.60648    0.91984  -2.834   0.0086 **
## hp          -0.02495    0.01365  -1.828   0.0786 .
## cyl         -0.74516    0.58279  -1.279   0.2119
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.509 on 27 degrees of freedom
## Multiple R-squared:  0.849,  Adjusted R-squared:  0.8267
## F-statistic: 37.96 on 4 and 27 DF,  p-value: 1.025e-10
```

Multivariable model explains almost 93% of variance, which is significant improvement from the 1st model.

4.ANOVA TEST

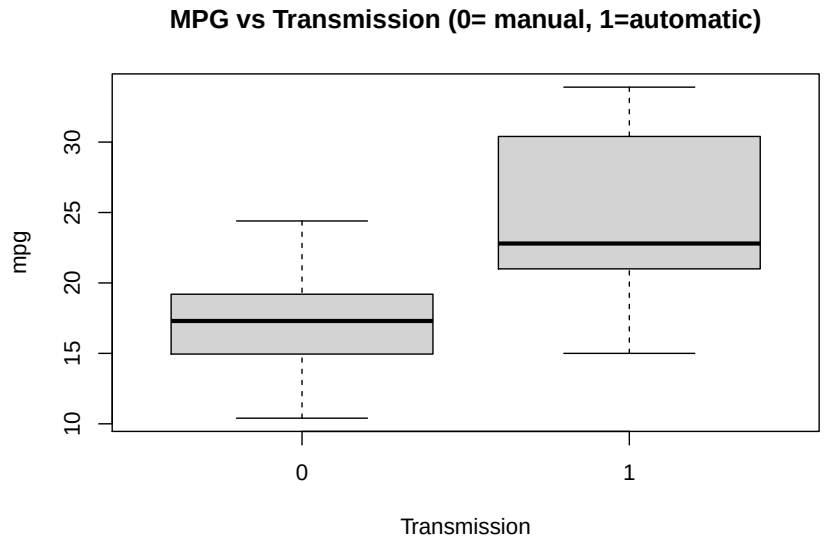
```
anova(fit5, fit2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt + hp + cyl
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.9
## 2     27 170.0  3     550.9 29.166 1.274e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The result of the anove test confirms significance of the multivariance model and reject the uni-variable model where only am is considered and rest of the variables are omitted.

Appendix

```
plot(mtcarsf$mpg~mtcarsf$am, main='MPG vs Transmission (0= manual, 1=automatic)', xlab='Transmission', y
```




```
par(mfrow = c(2,2))
plot(fit2)
```