Regression_Models_Assignment

Me

26/05/2021

Executive Summary

This assignment will study the relationships of variables in the mtcars dataset and find out whether automatic or manual transmission is better for MPG.

To quickly summarize the conclusion from our analysis here: manual transmission is better than automatic for MPG

Contents:

A short exploratory analysis before model fitting and model selection

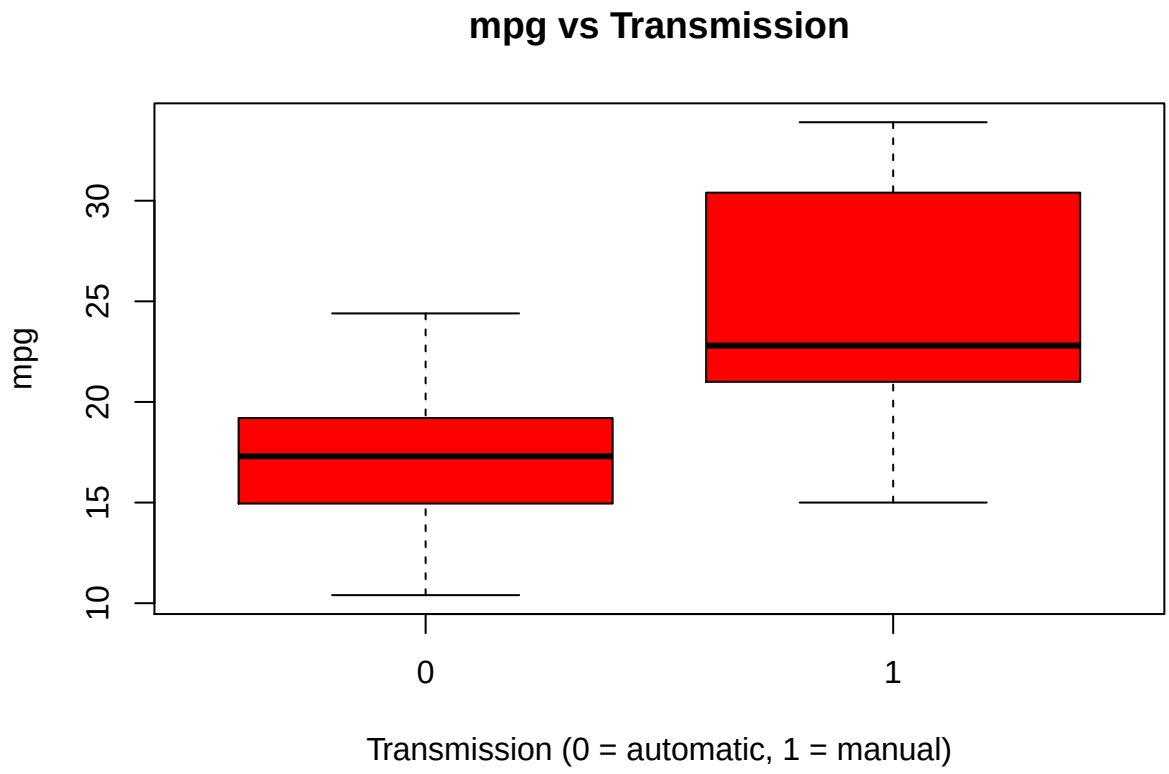
```
knitr::opts_chunk$set(echo = TRUE)

cars<- mtcars #copying mtcars dataset into a fresh new variable 'cars' as we don't want to disturb the
str(cars)
```

```
## '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 ...
```

```
head(cars)
```

```
##                   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
```

```
#plotting MPG on y axis and Transmission on x axis
boxplot(mpg~am, data= cars, col='red', xlab='Transmission (0 = automatic, 1 = manual)', main='mpg vs Tra
```



So as we can see, a quick boxplot analysis (without considering the influence of the other variables) is showing that manual transmisson is better than automatic transmission for MPG.

Main Report/Analysis

Here we have been asked to study the relationship between MPG and Transmission (automatic or manual)

So our outcome variable is mpg and the chief predictor variable will be am (0= automatic, 1= manual)

Model A (without adjusting for the rest of the variables)

Let's consider a model with only mpg (outcome var) and am(predictor) without adjusting for the other variables. We will be removing the intercept as we want to see the true slope coefficients

```
modelA<- lm(mpg~factor(am)-1, data = cars)
summary(modelA)
```

```
##
## Call:
## lm(formula = mpg ~ factor(am) - 1, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## factor(am)0   17.147      1.125   15.25 1.13e-15 ***
## factor(am)1   24.392      1.360   17.94  < 2e-16 ***
## ---
## 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:  0.9487, Adjusted R-squared:  0.9452
## F-statistic: 277.2 on 2 and 30 DF,  p-value: < 2.2e-16
```

So as we can see, slope coefficient of manual transmission is higher than that of automatic transmission. This means mpg increases more for every new/added car with manual transmission than one with automatic transmission. The standard errors are also reasonably low indicating less uncertainty in our estimates.

So this model suggests manual tranmission is better than automatic transmission

Model B (multi-variable model i.e. adjusting for the other important/necessary variables)

Since we didn't adjust for the effects of the other variables in model A, so we shouldn't take it's output very seriously. Hence, we should create a model that adjusts for the effects of the other variables.

However, we don't want any unnecessary variables in our model. So let's first create a nested model and test whether each variable is necessary with the help of the anova function.

Our alpha level will be 5% i.e. if the p value is less than 0.05, we reject the variable in question from the model as they are not necessary

```
#doing nested models now to find unnecessary variables

m1<- lm(mpg~factor(am), data = cars)
```

```
m2<- update(m1, mpg~factor(am)+factor(cyl))
m3<- update(m2, mpg~factor(am)+factor(cyl)+disp)
m4<- update(m3, mpg~factor(am)+factor(cyl)+disp+hp)
m5<- update(m4, mpg~factor(am)+factor(cyl)+disp+hp+drat)
m6<- update(m5, mpg~factor(am)+factor(cyl)+disp+hp+drat+wt)
m7<- update(m6, mpg~factor(am)+factor(cyl)+disp+hp+drat+wt+qsec)
m8<- update(m7, mpg~factor(am)+factor(cyl)+disp+hp+drat+wt+qsec+factor(vs))
m9<- update(m8, mpg~factor(am)+factor(cyl)+disp+hp+drat+wt+qsec+factor(vs)+factor(gear))
m10<- update(m9, mpg~factor(am)+factor(cyl)+disp+hp+drat+wt+qsec+factor(vs)+factor(gear)+carb)
anova(m1, m2, m3, m4, m5, m6, m7, m8, m9, m10)
```

```
## Analysis of Variance Table
##
## Model  1: mpg ~ factor(am)
## Model  2: mpg ~ factor(am) + factor(cyl)
## Model  3: mpg ~ factor(am) + factor(cyl) + disp
## Model  4: mpg ~ factor(am) + factor(cyl) + disp + hp
## Model  5: mpg ~ factor(am) + factor(cyl) + disp + hp + drat
## Model  6: mpg ~ factor(am) + factor(cyl) + disp + hp + drat + wt
## Model  7: mpg ~ factor(am) + factor(cyl) + disp + hp + drat + wt + qsec
## Model  8: mpg ~ factor(am) + factor(cyl) + disp + hp + drat + wt + qsec +
##     factor(vs)
## Model  9: mpg ~ factor(am) + factor(cyl) + disp + hp + drat + wt + qsec +
##     factor(vs) + factor(gear)
## Model 10: mpg ~ factor(am) + factor(cyl) + disp + hp + drat + wt + qsec +
##     factor(vs) + factor(gear) + carb
##    Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1      30 720.90
## 2      28 264.50  2    456.40 33.3392 6.108e-07 ***
## 3      27 230.46  1     34.04  4.9725   0.03801 *
## 4      26 183.04  1     47.42  6.9280   0.01642 *
## 5      25 182.38  1      0.66  0.0961   0.75989
## 6      24 150.10  1     32.28  4.7161   0.04275 *
## 7      23 141.21  1      8.89  1.2995   0.26848
## 8      22 139.02  1      2.18  0.3189   0.57887
## 9      20 134.00  2      5.02  0.3668   0.69774
## 10     19 130.05  1      3.95  0.5771   0.45677
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On basis of this above output from the anova function and setting our alpha level to 5% or 0.05, we can conclude the variables drat, qsec, vs, gear and carb are unnecessary for our model.

Therefore, the necessary variables/predictors for our model are: am, cyl, disp, hp, wt so we will create our model with these 5 chosen predictors

Creating our final model i.e. Model B

```
modelB<- lm(mpg~factor(am)+factor(cyl)+disp+hp+wt-1, data= cars)
summary(modelB)
```

```
##
## Call:
```

```
## lm(formula = mpg ~ factor(am) + factor(cyl) + disp + hp + wt -
##     1, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9374 -1.3347 -0.3903  1.1910  5.0757
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## factor(am)0  33.864276   2.695416  12.564 2.67e-12 ***
## factor(am)1  35.670376   2.122096  16.809 3.90e-15 ***
## factor(cyl)6 -3.136067   1.469090  -2.135   0.0428 *
## factor(cyl)8 -2.717781   2.898149  -0.938   0.3573
## disp          0.004088   0.012767   0.320   0.7515
## hp           -0.032480   0.013983  -2.323   0.0286 *
## wt           -2.738695   1.175978  -2.329   0.0282 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.453 on 25 degrees of freedom
## Multiple R-squared:  0.9893, Adjusted R-squared:  0.9863
## F-statistic: 329.9 on 7 and 25 DF,  p-value: < 2.2e-16
```

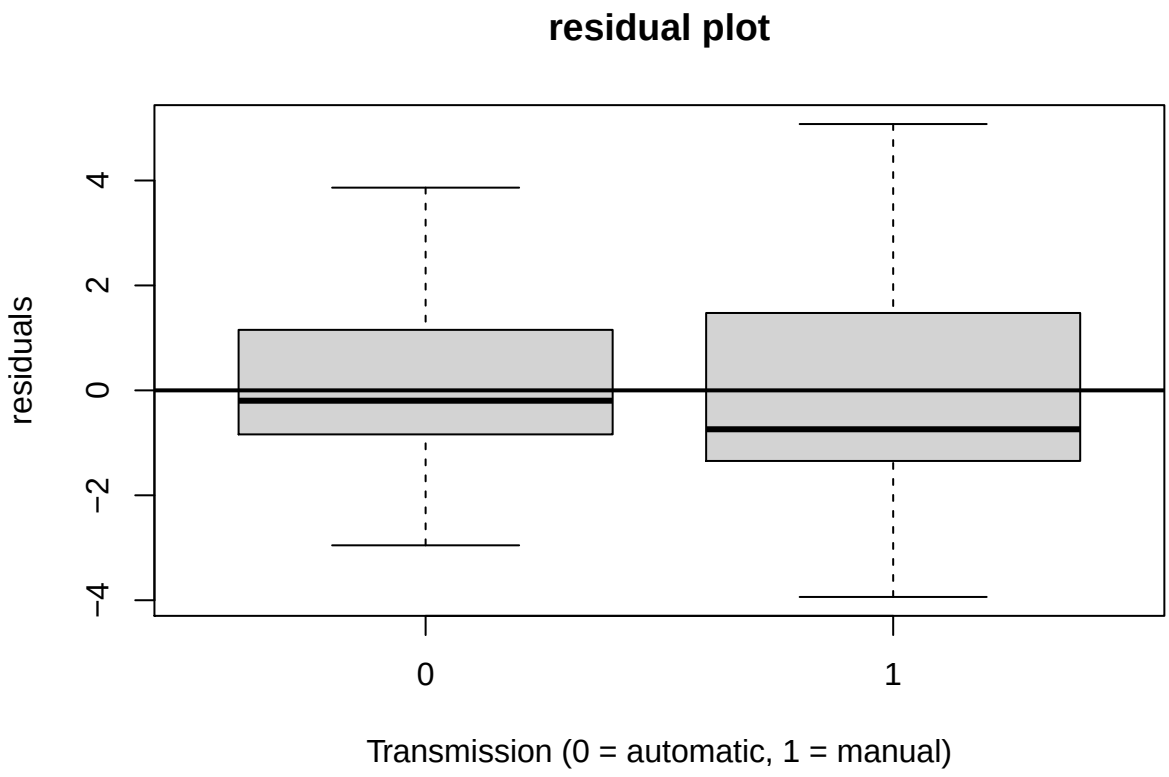
So as we can see, true slope coefficient for manual transmission=35.670376, is still larger than that of automatic transmission=33.864276. So mpg increases more for every new car with manual transmission than one with automatic transmission by 1.8061 miles per gallon (35.670376-33.864276) .

However, this time standard errors are over 2 which indicates much higher uncertainty in our estimates.

Conclusion: manual transmission is better for MPG than automatic transmission

Residual plot: using model B

```
e_modelB<- resid(modelB)
boxplot(e_modelB~cars$am,  xlab='Transmission (0 = automatic, 1 = manual)', ylab = 'residuals', main= '
abline(h=0, lwd=2)
```



Appendices

Here we create a box plot for showing the relationship between our estimated mpg values and am (0= automatic, 1= manual) from modelB (adjusting for all of the other important/necessary variables)

```
boxplot(predict(modelB)~cars$am, col='red', xlab='Transmission (0 = automatic, 1 = manual)', ylab='predi
```