Regresion Models Project

Abdelrahman Elsehaily

October 7, 2017

1.Executive Summary

This is the project of the Regression models course, the goal of the project is to explore the relationship between a set of variables and miles per gallon (MPG) (outcome) in the motor cars dataset, we are particularly interested in the following two questions:
"Is an automatic or manual transmission better for MPG"
"Quantify the MPG difference between automatic and manual transmissions"
First I used t-test to see if there is significant difference between manual and automatic then used linear regression to quntify this difference then adjusted that model to enhance the **r-squared** and used **anova** to compare the different model.

2.Analysis

Is the difference between maual and automatice is significant?

As seen in figure 2 automatic has mpg lower values than manual so we want to check if that difference is significant.

```
mtcars$am<-ifelse(mtcars$am==0,"automatic","manual")
t.test(mpg~am,data = mtcars)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by 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 automatic    mean in group manual
##                17.14737                24.39231
```

The p-value is lower than 0.0025 (tow-sided test) so there is a significant difference in the mean of the two groups.

Quntifying the difference between the manual and automatic

using am only implement linear regression model

```
amModel<-lm(mpg~ am,data=mtcars)
summary(amModel)
```

```
##
## 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 ***
## ammanual       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
```

from the summary autmatic cars has 17.147 mpg **on average** and manual cars is higher with value 7.245

Adjustment

The adjusted R-squared is 0.3385 which mean that am could only explain 33.8 % of the variance, so we should add another variables to our model but that should depend on the correlation between the mpg and the other variables (see appendix 2)

```
allCorModel<-lm(mpg~ am+wt+qsec,data=mtcars)
allVarModel<-lm(mpg~ .,data=mtcars)
anova(amModel,allCorModel,allVarModel)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt + qsec
## Model 3: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 169.29  2    551.61 39.2687 8.025e-08 ***
## 3     21 147.49  7     21.79  0.4432    0.8636
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Residual Sum Squared (RSS) has decreased from 720 to 169.29 in the second one and as **p-value<0.05** this model is **significantly** better than the one with the variable am only.
In the last model the Residual saured error has decreased but the p-value is bigger than 0.05 so this model is not significantly better than the prevoius so I think using all variables may have multicolinearity problem represent an overfitted model.

Diagnostics

checking the Varience inflation factor to see how the variable affect each other in the last two model

```
library(car)
vif(allCorModel)
```

```
##       am       wt     qsec
## 2.541437 2.482952 1.364339
```

```
vif(allVarModel)
```

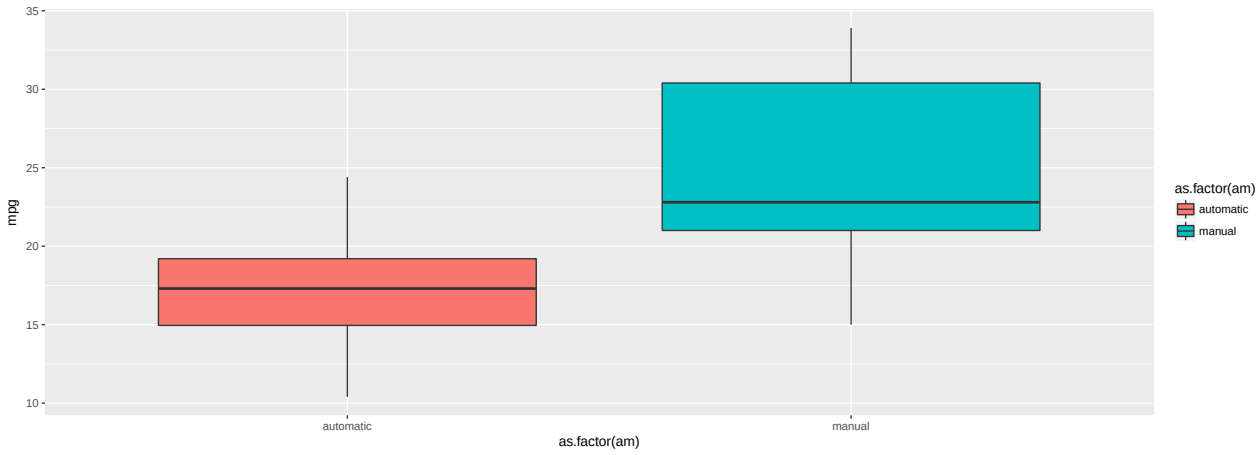
```
##       cyl      disp        hp      drat        wt      qsec        vs
## 15.373833 21.620241  9.832037  3.374620 15.164887  7.527958  4.965873
##        am      gear      carb
##  4.648487  5.357452  7.908747
```

so you see how large the vif in the last model that because of the multicolinearity issue also our chosen model has vif bigger than 2 but it is relatively low compared to the last model.

Appendix

Appendix 1:The difference in mpg with manual and automatice

```
library(ggplot2)
ggplot(data = mtcars,aes(x=as.factor(am),y=mpg,fill=as.factor(am)))+geom_boxplot()
```

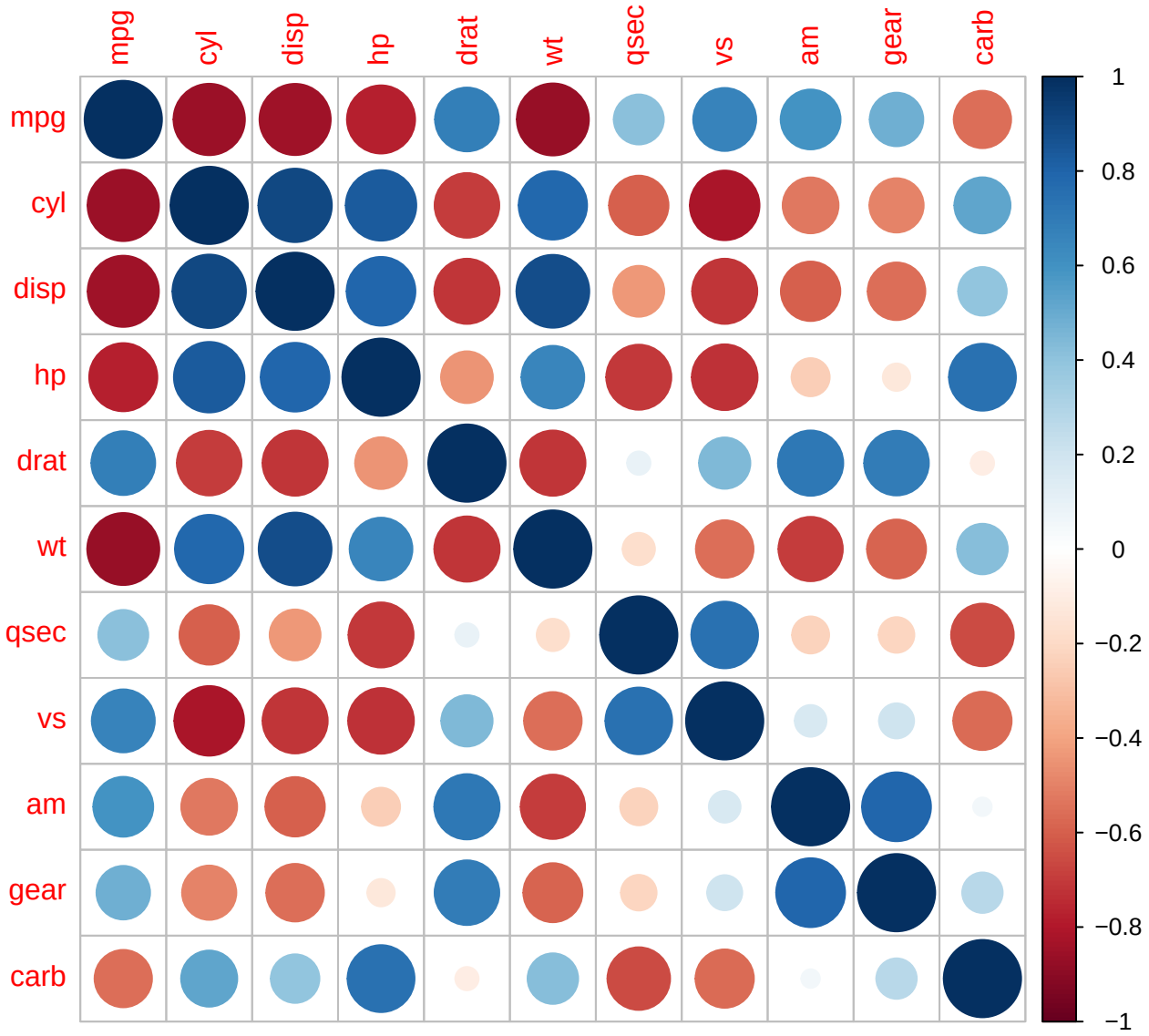


Appendix 2: correlation between the mtcars columns

```
data("mtcars")
library(corrplot)
```

```
## corrplot 0.84 loaded
```

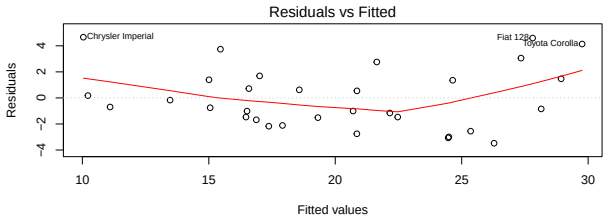
```
corrplot(cor(mtcars[, ]))
```

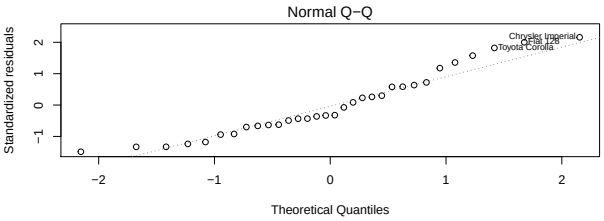


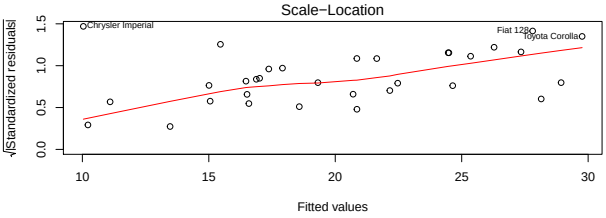
Appendix 3:The dignostic plots of the selected model

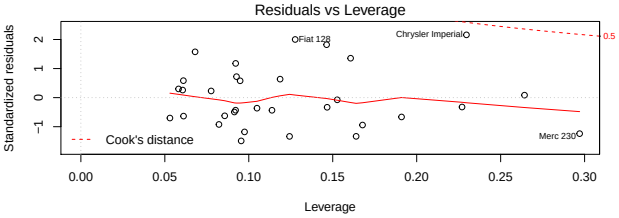
seems from the plot that there's not strong pattern in the residuals

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


Residuals vs Fitted
Chrysler Imperial
Fiat 128
Toyota Corolla
Residuals
Fitted values


Normal Q−Q
Chrysler Imperial
Fiat 128
Toyota Corolla
Standardized residuals
Theoretical Quantiles


Scale−Location
Chrysler Imperial
Fiat 128
Toyota Corolla
√|Standardized residuals|
Fitted values


Residuals vs Leverage
Chrysler Imperial
Fiat 128
Merc 230
Cook's distance
0.5
Standardized residuals
Leverage