RegresionModelsProject

Abdelrahman Elsehaily

October 3, 2017

1.Introduction

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.

2. Analysis

2.1.Reformatting the data

After reading the documentaion I think this column shoud be factorss

```
data("mtcars")
carsdata=mtcars
#assign character values to am column so it doesn't affect the regression models and be more readable
carsdata$am<-ifelse(mtcars$am==1,"manual","auto")
carsdata$gear<-as.factor(carsdata$gear)
carsdata$cyl<-as.factor(carsdata$cyl)
```

2.2.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.

```
t.test(mpg~am,data = carsdata)
```

```
##
##  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 auto 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. ### 2.3.Modeling First using am only implement linear regression model

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

```
##
## Call:
## lm(formula = mpg ~ am, data = carsdata)
##
```

```
## 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
```

The adjuster R-squared is 0.3385 which mean that am could only explain 33.8 % of the variance

Nested Models

It is obvious that am only is not a good way to predict the mpg, so we should add another variables to our model but that should depend on the correlation between the mpg and the other variables (appendix 2)

Comparing Models

The weight varible has the highest (negative) correlation with mpg(figure 3) so we gonna add this variable to our model and that will be our first model.
The second will use all the variables that is highly correlated to mpg.
The third will use all the variables in our dataset.

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

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + cyl + disp + wt
## Model 4: 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     29 278.32  1    442.58 63.0133 9.325e-08 ***
## 3     27 188.43  2     89.89  6.3995  0.006758 **
## 4     21 147.49  6     40.93  0.9713  0.468405
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Residual Sum Squared (RSS) has decreased from 720 to 278 in the first one and as **p-value<0.05** this model is **significantly** better than the one with the variable am only.
In the second model the Residual saured error has decreased from 278.3 to 188.43, the p-value is lower than 0.05 so this model is significantly better than the one with the weight,in the third model the residuals has decreases but the p-value is **not significant** at all so I think using all variables may represent an overfitted model.

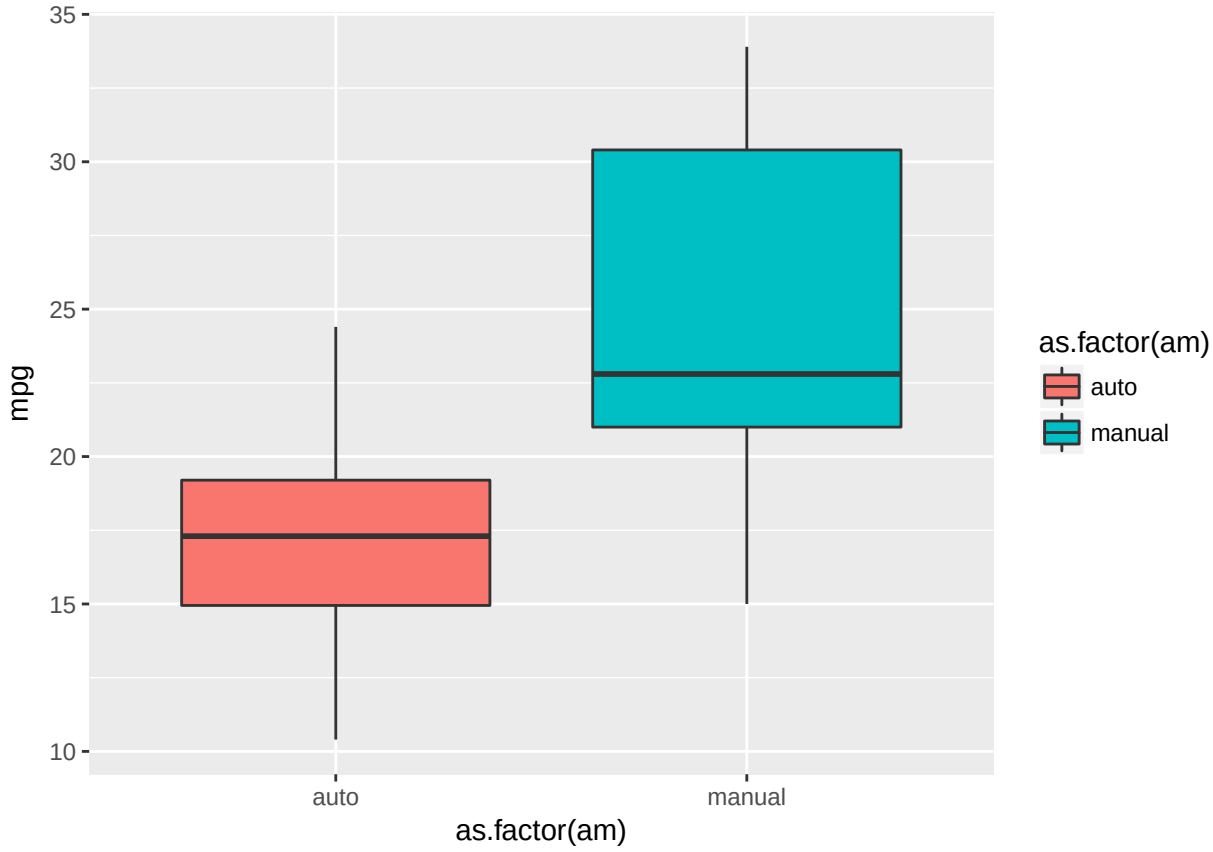
Appendix

Appendix 1

```
library(ggplot2)
```

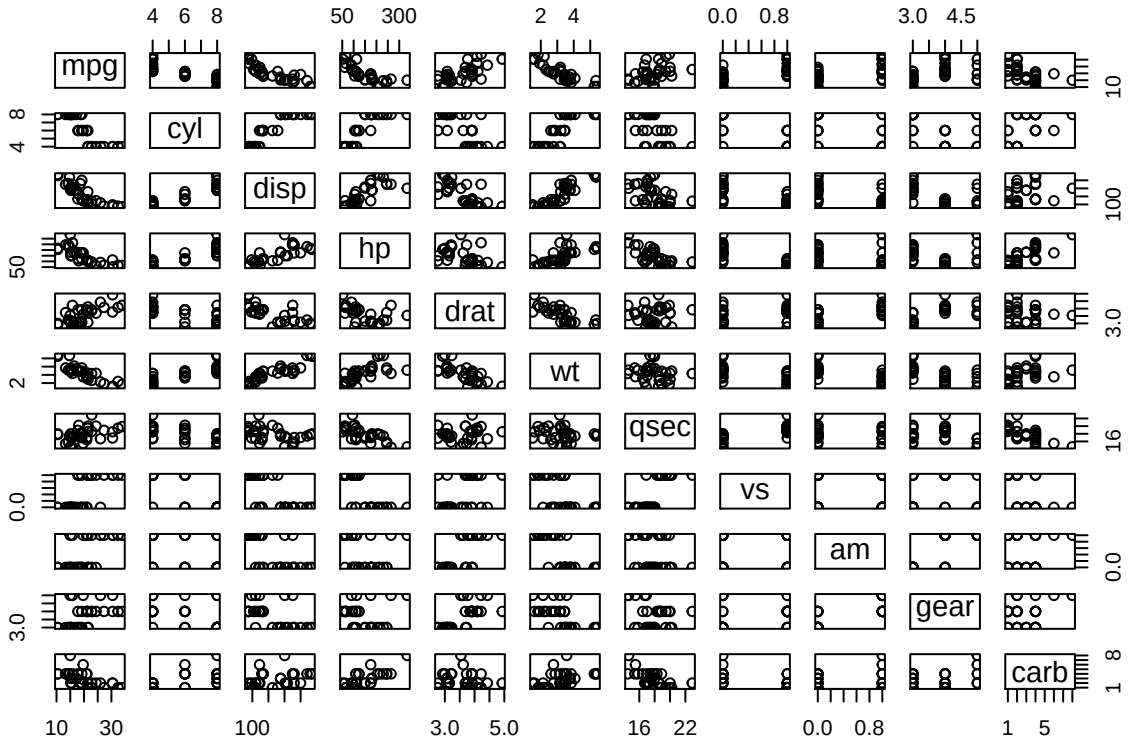
```
## Warning: package 'ggplot2' was built under R version 3.4.1
```

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



Appendix 2

```
pairs(mtcars)
```



Correlation

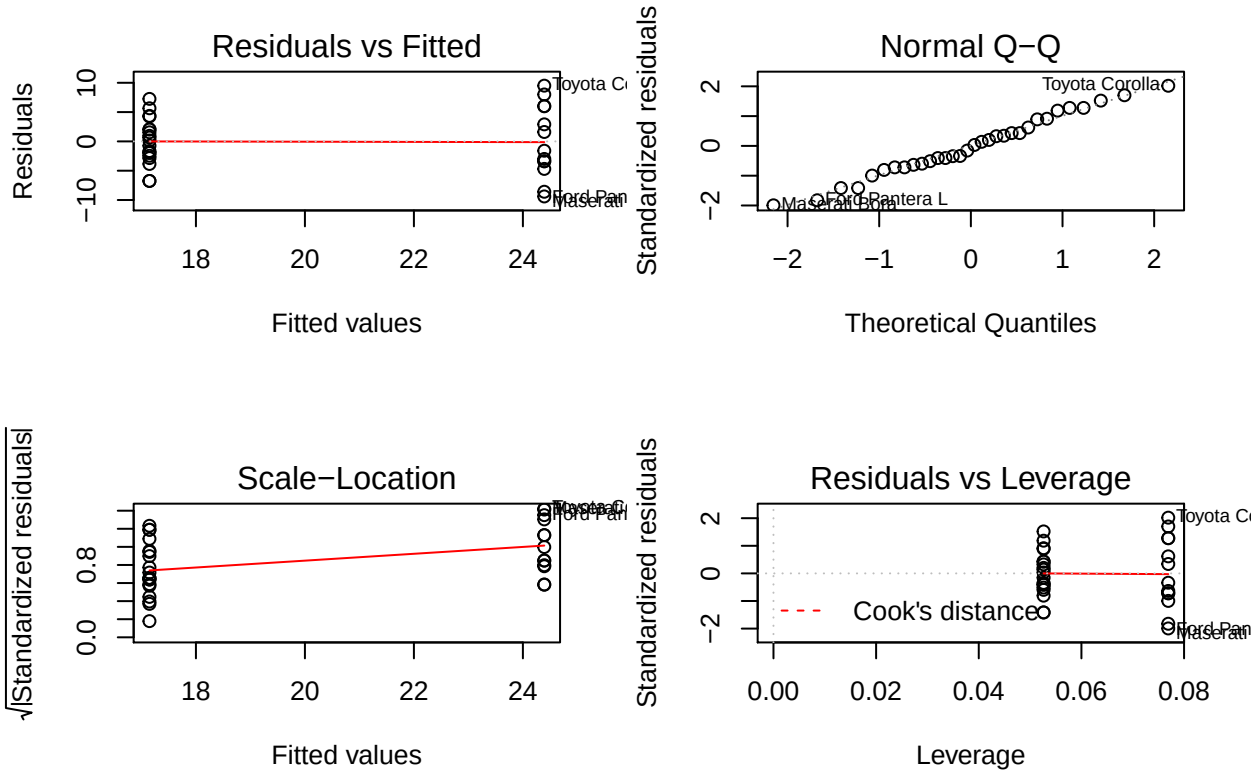
```
cor(mtcars$mpg,mtcars)
```

```
##      mpg       cyl       disp         hp      drat         wt     qsec
## [1,]   1 -0.852162 -0.8475514 -0.7761684 0.6811719 -0.8676594 0.418684
##             vs        am      gear       carb
## [1,] 0.6640389 0.5998324 0.4802848 -0.5509251
```

Appendix 3

The model with am only.

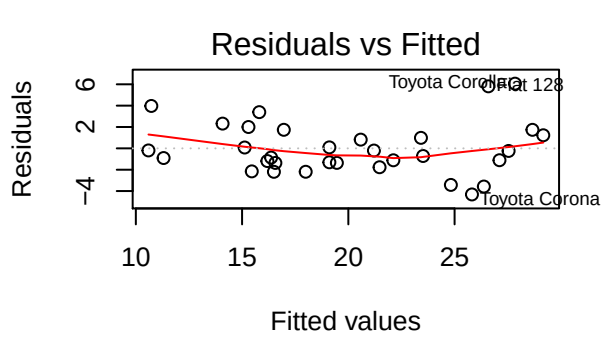
```
par(mfrow=c(2,2))
plot(amModel)
```

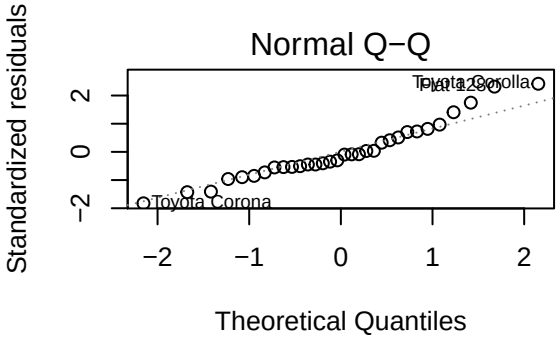


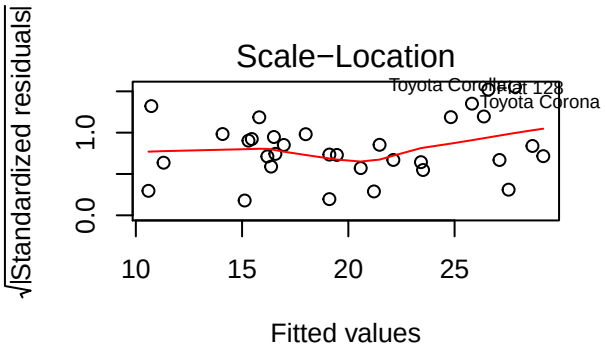
Appendix 4

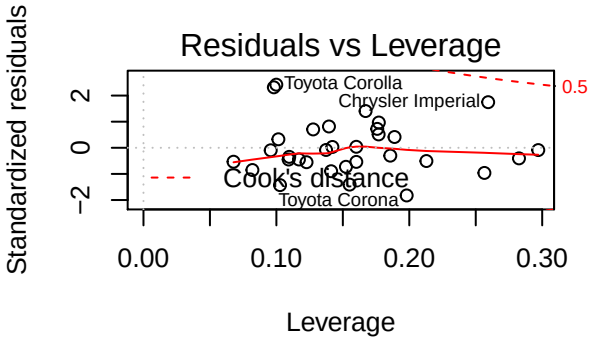
The model with the highly correlated variables

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


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Toyota Corona


Normal Q−Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Fiat 128
Toyota Corona


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


Residuals vs Leverage
Standardized residuals
Leverage
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
0.5