Reg_Models Course Project

Sai Rahul Ponnana

14 August 2020

Executive Summary

In this study we look at the cars dataset comprising of different aspects of automobile design for 32 automobiles, to explore the relationship between these aspects with the miles per gallon. We specifically focus on the following two questions being is an automatic or manual transmission better for MPG and how to quantify this MPG difference between automatic and manual transmissions.

Data Preprocessing

First, we change the 'am' variable of the dataset which denotes if a car is automatic or manual transmission to a factor variable. We also other variables factor just as to make them discrete instead of continuous.

```
data(mtcars)
names(mtcars)
```

```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"
"am"   "gear"
## [11] "carb"
```

Analysis

As we can see, there are 11 variables in the dataset. We are interested in the relationship between mpg and other variables, so first we check the correlation between mpg and other variables by using the cor() function.

```
cor(mtcars$mpg,mtcars[,-1])
```

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

Exploratory Analysis

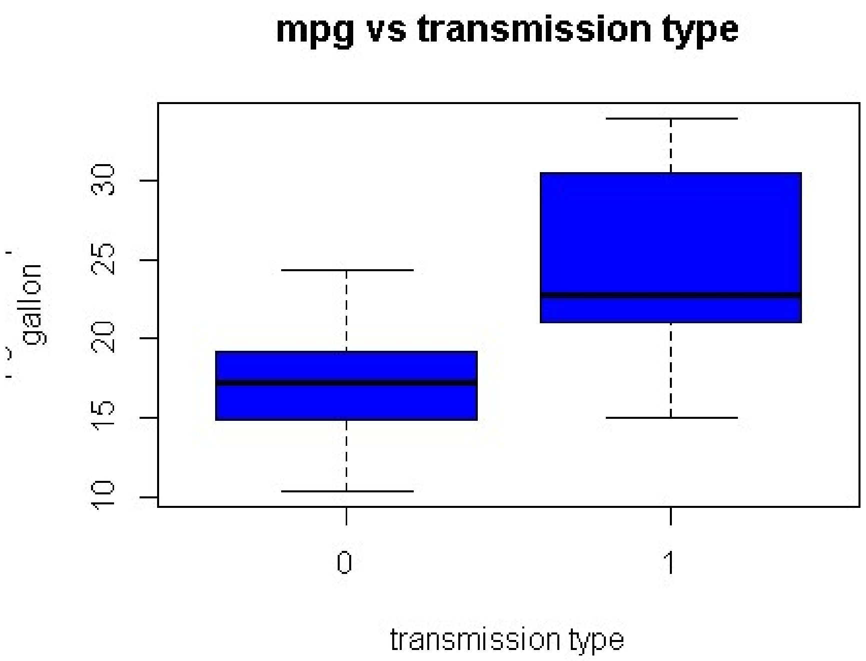
First let's take a look at the dataset itself to know about the fields it contains.

```
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 ...
```

To see the relationship between the mpg and am more clearly lets create a boxplot.

```
library(ggplot2)
boxplot(mtcars$mpg ~ mtcars$am, data = mtcars, outpch = 19, 
ylab="mpg:miles per 
gallon",xlab="transmission type",main="mpg vs transmission type", 
col="blue")
```



The plot clearly shows that cars with manual transmission do have higher mpg as compared to the one's with automatic transmission. However there might be other factors which we might be overlooking. Hence before creating a model we should look at other parameters which have high correlation with the variable. Lets look at all the variables whose correlation with mpg is higher than the am variable.

Model Selection

Now that we know mpg variable has stronger correlations with other variables too apart from just am, we can't base our model solely on this one variable as it will not be the most accurate one. Let's start this process by fitting mpg with just am.

```
first <- lm(mpg ~ am, mtcars)
summary(first)
```

```
## 
## 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
```

In this case p-value is quite low but the R-squared value is the real problem. Hence, let's now go to the other extreme end and fit all variables with mpg.

```
last <- lm(mpg ~ ., mtcars)
summary(last)
```

```
## 
## Call:
## lm(formula = mpg ~ ., data = mtcars)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.4506 -1.6044 -0.1196  1.2193  4.6271 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

Here R-squared values have definitely improved but the p-value becomes the problem now which is caused most probably due to overfitting. So, lets use ‘step’ method to iterate over the variables and obtain the best model.

```
best <- step(last, direction = "both", trace = FALSE)
summary(best)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am            2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

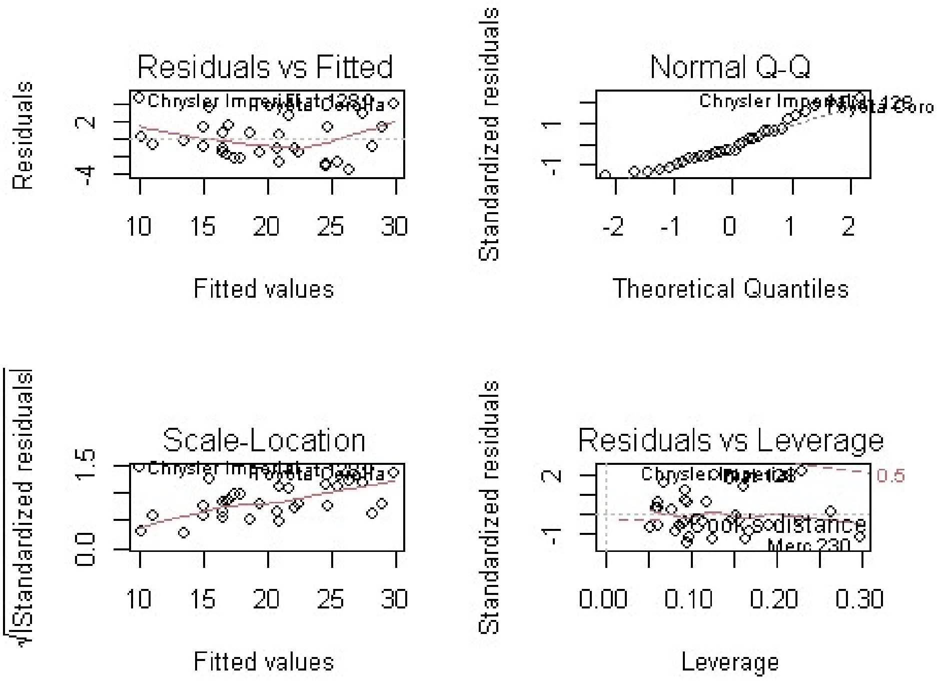
Here the R-squared value is pretty good and also p-values are quite significant. Hence undoubtedly this is the best fit for us.

Model Examination

Residual check and Diagnostics plot

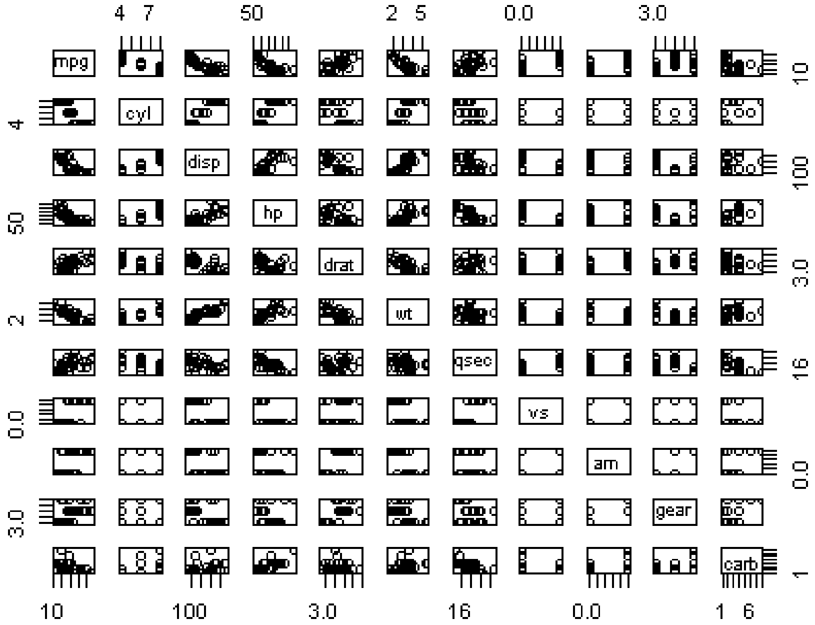
The best model we obtained i.e., ‘best’ depicts the dependance of mpg over wt and qsec other than am. Let’s plot and study some residual plots to understand more about the ‘best’ fit.

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



Scatterplots

```
pairs(mpg ~ ., data = mtcars)
```



Conclusion

The first question whether automatic or manual is better for mpg can be answered using all the models created as holding all the other parameters constant, manual transmission increases the mpg.

However the second question is a little difficult to answer. Based on 'best' fit model, we conclude that cars with manual transmission have 2.93 more mpg than that of automatic with $p < 0.05$ and R-squared 0.85.

Residuals vs Fitted plot however shows something is missing from the model which might be a problem due to a small sample size which is 32 observations. Even though the conclusion that manual has better performance with respect to mpg, whether the model will git all future observations will be doubtful.