Regression Models - mtcars data

Executive Summary

In this study, we are going to look at the data set "mtcars" and find what variables affect the efficiency value of miles per gallon (MPG). Then, we decide which one is more efficient in between automatic and manual transmission as well as quantifying their differences.

To achieve our objectives, we will perform the procedures below.
- Data preprocessing
- Exploratory Analysis
- Model Selection
- Model Examination

Data pre processing

```
# load data
data("mtcars")
mtcars$vs <- factor(mtcars$vs)
mtcars$am.label <- factor(mtcars$am, labels=c("Automatic","Manual"))
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb  am.label
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4    Manual
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4    Manual
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1    Manual
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1 Automatic
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2 Automatic
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1 Automatic
```

Exploratory Analysis

```
summary(mtcars)
```

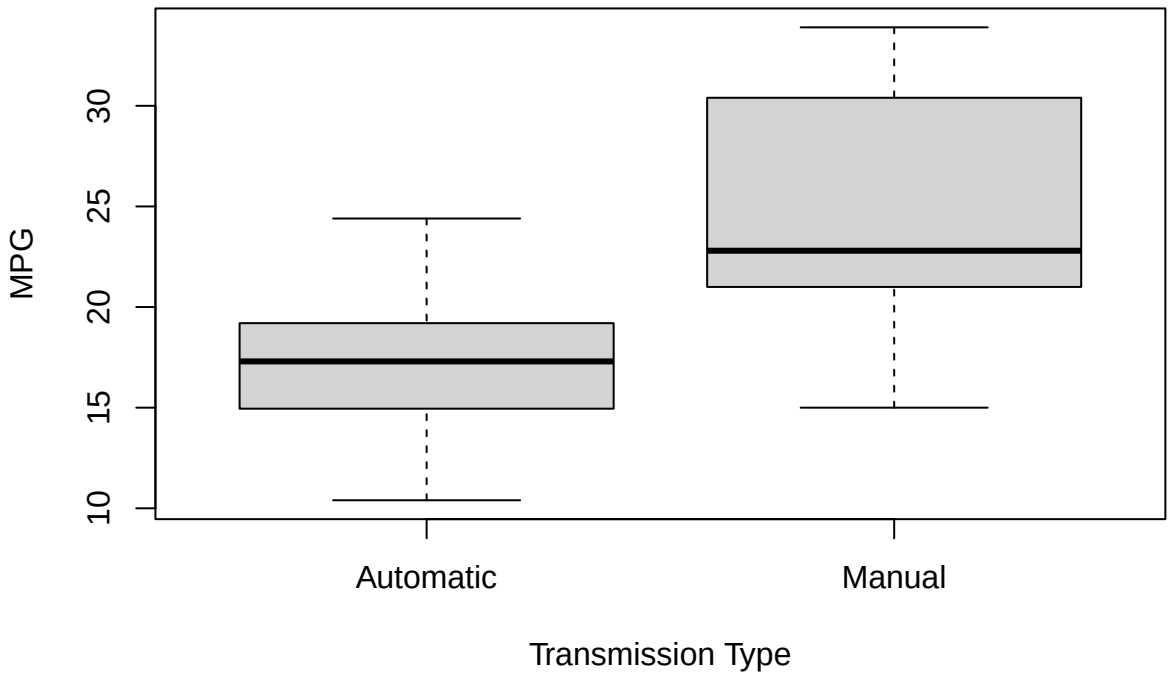
```
##       mpg             cyl             disp             hp
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0
```

```
##       drat             wt             qsec       vs          am          gear
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   0:18   Min.   :0.0000   3:15
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1:14   1st Qu.:0.0000   4:12
##  Median :3.695   Median :3.325   Median :17.71          Median :0.0000   5: 5
##  Mean   :3.597   Mean   :3.217   Mean   :17.85          Mean   :0.4062
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90          3rd Qu.:1.0000
##  Max.   :4.930   Max.   :5.424   Max.   :22.90          Max.   :1.0000
##  carb        am.label
##  1: 7   Automatic:19
##  2:10   Manual   :13
##  3: 3
##  4:10
##  6: 1
##  8: 1
```

```
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb  am.label
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4    Manual
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4    Manual
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1    Manual
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1 Automatic
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2 Automatic
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1 Automatic
```

```
boxplot(mpg ~ am.label, data = mtcars, ylab = "MPG", xlab = "Transmission Type")
```



0 = Automatic Transmission 1 = Manual Transmission

Without considering other factors, as shown on the plot above, the manual transmission type shows better efficiency since it has higher MPG value. To dig it deeper, we actually can see the degree of correlation of all variables toward MPG.

```
data("mtcars")
correlation <- cor(mtcars$mpg, mtcars)
correlation
```

```
##      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
```

Model Selection

After checking the correlation values, we can judge that this model cannot rely only one variable since it will be inaccurate. We can try by initially fit mpg into am only.

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

```
##
## 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
```

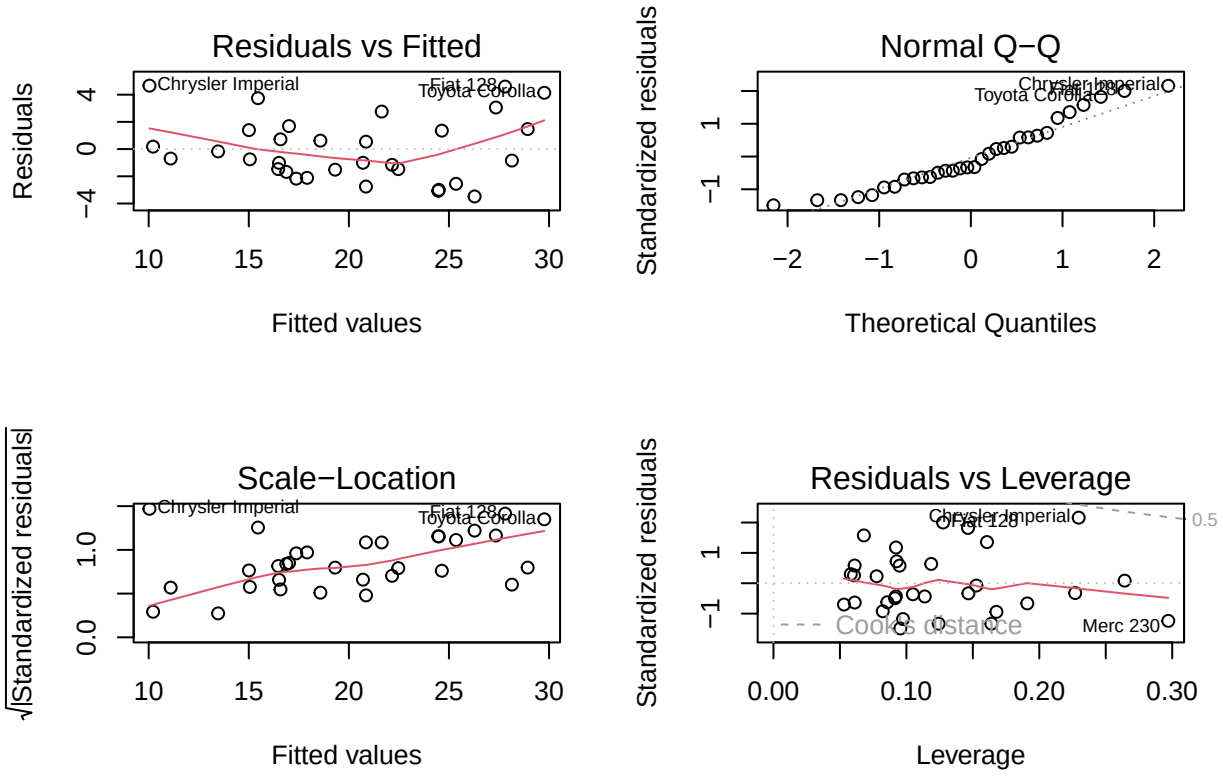
```r
best <- step(fit, direction = "both", trace = FALSE)
summary(best)
```

```
##
## 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
```

p value is increased but, it might be an over fit, so in second try, we are looking into finding the best model which will be used.

Model Examination

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



It shows that selected model has a good fit and good correlation is depicted on the graph Normal Q-Q.

Conclusions

After conducting the analysis it can be concluded that manual is better than automatic transmission in terms of "miles per gallon" value. So the answer for the first question is no. However, for second question, it is still difficult to quantify. From the model, we can see that manual one will have an average of 2.9358 higher miles/gallon than the automatic car with $p < 0.05$ and $R\hat{}2 = 0.85$. Perhaps we need more sample data to improve the confidence and accuracy.