Motor Cars

MT

12/18/2020

Executive Summary

The report will include analysis of car transmission and estimate the relationship between each of the variables from mtcars dataset. Using regression models, prediction models were also generated to help determine which is the best fit for the dataset collected.

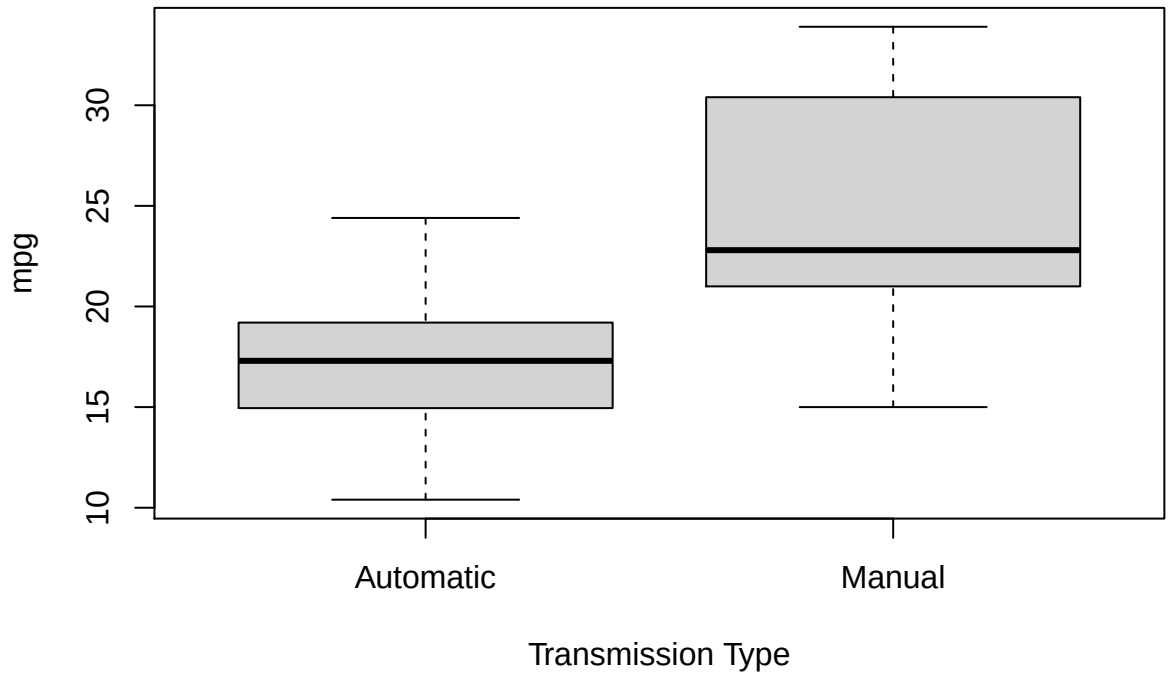
Exploring Dataset

```
library(datasets)
data(mtcars)
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
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375
##  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
##        am              gear            carb
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
##  Median :0.0000   Median :4.000   Median :2.000
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

```
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
```

```r
mtcars$carb <- factor(mtcars$carb)
mtcars$am <- factor(mtcars$am,labels=c('Automatic','Manual'))
boxplot(mpg~am, mtcars, xlab="Transmission Type")
```



Based on the boxplot, it can be observes that there might be a significant difference in MPG for the different transmission.

#Simple linear Regression Model

```r
mpgam<- lm(mpg~factor(am), mtcars)
summary(mpgam)
```

```
##
## Call:
## lm(formula = mpg ~ factor(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 ***
## factor(am)Manual    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
```

Looking at the single variate model, the R^2 values were only 0.3598. Which also means that only 35.9% of variance can be explained by this model and does not provide a good fit to the dataset observed. Hence, other predictors are to be further assessed for any influence to create a better fit of model.

#Multivariate Regression Model

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

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5087 -1.3584 -0.0948  0.7745  4.6251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.87913   20.06582   1.190   0.2525
## cyl6        -2.64870    3.04089  -0.871   0.3975
## cyl8        -0.33616    7.15954  -0.047   0.9632
## disp         0.03555    0.03190   1.114   0.2827
## hp          -0.07051    0.03943  -1.788   0.0939 .
## drat         1.18283    2.48348   0.476   0.6407
## wt          -4.52978    2.53875  -1.784   0.0946 .
## qsec         0.36784    0.93540   0.393   0.6997
## vs1          1.93085    2.87126   0.672   0.5115
## amManual     1.21212    3.21355   0.377   0.7113
## gear4        1.11435    3.79952   0.293   0.7733
## gear5        2.52840    3.73636   0.677   0.5089
## carb2       -0.97935    2.31797  -0.423   0.6787
## carb3        2.99964    4.29355   0.699   0.4955
## carb4        1.09142    4.44962   0.245   0.8096
## carb6        4.47757    6.38406   0.701   0.4938
## carb8        7.25041    8.36057   0.867   0.3995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.833 on 15 degrees of freedom
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.779
## F-statistic:  7.83 on 16 and 15 DF,  p-value: 0.000124
```

```
mpgam_cor<- lm(mpg~am+cyl+disp+hp+wt, mtcars)
mpgam_cor_nowt<- lm(mpg~am+cyl+hp+disp, mtcars)
mpgam_cor_noam<- lm(mpg~cyl+hp+disp, mtcars)
mpgam_cor_nodisp<- lm(mpg~am+cyl+hp+wt, data=mtcars)
mpgam_cor_noamdisp<- lm(mpg~cyl+hp+wt, data=mtcars)
```

```r
anova(mpgam_all,
      mpgam_cor,
      mpgam_cor_nowt,
      mpgam_cor_noam,
      mpgam_cor_nodisp,
      mpgam_cor_noamdisp)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## Model 2: mpg ~ am + cyl + disp + hp + wt
## Model 3: mpg ~ am + cyl + hp + disp
## Model 4: mpg ~ cyl + hp + disp
## Model 5: mpg ~ am + cyl + hp + wt
## Model 6: mpg ~ cyl + hp + wt
##   Res.Df    RSS  Df Sum of Sq      F   Pr(>F)
## 1     15 120.40
## 2     25 150.41 -10   -30.006 0.3738 0.939655
## 3     26 183.04  -1   -32.630 4.0652 0.062050 .
## 4     27 225.12  -1   -42.078 5.2422 0.036960 *
## 5     26 151.03   1    74.092 9.2305 0.008301 **
## 6     27 160.78  -1    -9.752 1.2149 0.287730
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at the models fitted and compared with anova analysis, it shows that while there are correlations. Model fitted with `cyl`, `hp`, `wt`, `am` would fit the dataset better with the lowest RSS. With the p-value below 0.05, we cannot reject the null hypothesis that `cyl`, `hp` and `wt` does not have an influence. Using stepwise analysis also confirms the same final four variable to be the best variable to be included as the predictors of the model.

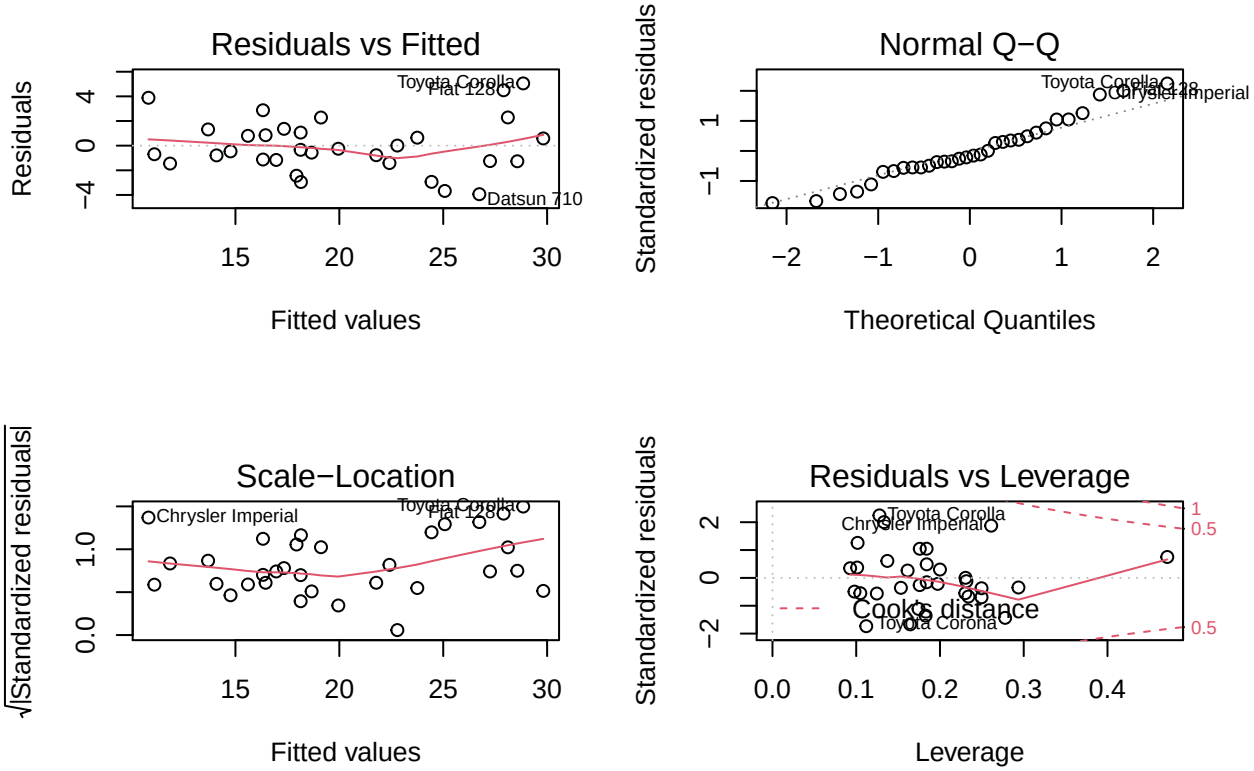
#Analysis and Conclusion ##Inference

```r
summary (mpgam_cor_nodisp)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + hp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## amManual     1.80921    1.39630   1.296  0.20646
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

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




```
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
```

Basing on the summary, the adjusted R^2 suggests that the predicted model is able to conclude 84% of the variability. T-Test also shows shows that automatic and manual are significantly different from each other.

The points in residuals vs.Fitted plots seem to be randomly scattered on showing that it is independent. The normal Q-Q plot consist of points which closely fits the line indicating that they are normal. The scale-location plot are scattered constantly, indicating constant variance. The leverage plots, showed that there are some points which may potentially have some strong influences. ## Leverage Diagnosis

```
lev<- hatvalues(mpgam_cor_nodisp)
tail(sort(lev),3)
```

```
##       Toyota Corona Lincoln Continental       Maserati Bora
##           0.2777872           0.2936819           0.4713671
```

```
tail(sort(dfbeta(mpgam_cor_nodisp)[,1]),3)
```

```
## Cadillac Fleetwood     Toyota Corolla         Volvo 142E
##          0.7309046          0.8005292          0.8247396
```

Based on the diagnostics, it seems like Toyota Corolla may portential have a strong influence to the model's prediction hence further assessment might be required to determine if it should be excluded.

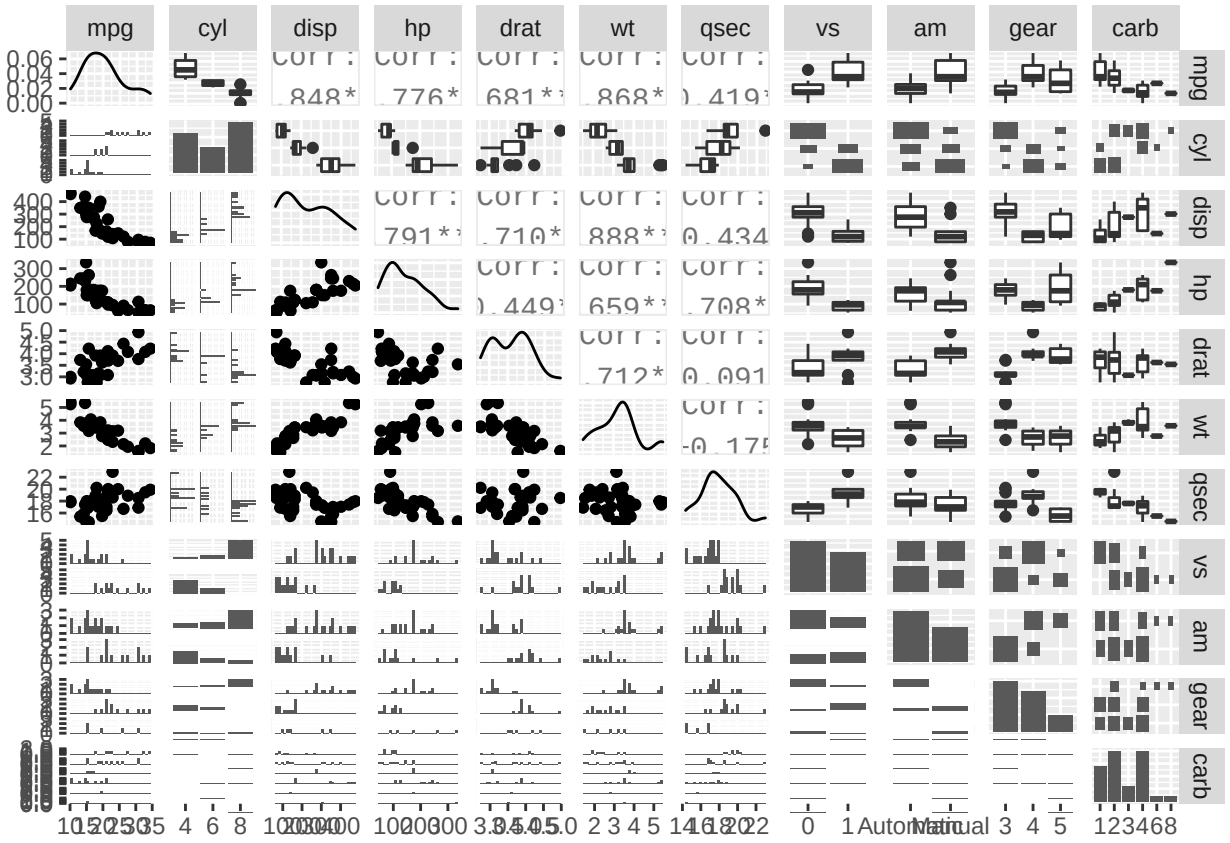
##Conclusion to Questions 1. "Is an automatic or manual transmission better for MPG"

Manual transmission is better for MPG, as observed from the data, manual transmissions increase the mileage of each vehicle, however other variables can also influence the difference in mileage within vehicles with manual transmission too.

2. "Quantify the MPG difference between automatic and manual transmissions" The mean difference between automatic and manual transmission is 1.8.

#Appendix

```
library(GGally)
library(ggplot2)
ggpairs(mtcars)
```



Plots showing the correlation between two variables, distribution and the scatterplots. The plot provides a quick exploratory analysis to identify the plots of interest.