Factors affecting Mile per Gallons of Motors

ThetLwinLwin

1/3/2021

Executive Summary

In this project we will explore the relationship between a set of variables and miles per gallon (MPG) (outcome). We are particularly interested in the following two questions: - 'is an automatic or manaual transmission better for MPG' - 'Qunatify the MPG difference between automatic and manual transmissions' The data we used is **mtcars** dataset which is readily available in R.The analyses will use the regression models to seek the answers. The report is composed of four main parts. 1. Exploratory Analysis 2. Regression Analysis 3. Residual Analysis 4. Conclusions

1. Exploratory Analysis

The basic observation to do is to know the dataset before anything has done. Meaning of each variable in dataset and what kind of variables are composed in that dataset is important to investigate. The code chunk below will show some useful information about the dataset.

- **mpg**: Miles/(US) gallon
- **cyl**: Number of cylinders
- **disp**: Displacement (cu.in.)
- **hp**: Gross horsepower
- **drat**: Rear axle ratio
- **wt**: Weight (lb/1000)
- **qsec**: 1/4 mile time
- **vs**: V/S
- **am**: Transmission (0 = automatic, 1 = manual)
- **gear**: Number of forward gears
- **carb**: Number of carburetors

The data has 11 variables with 32 observatrions. It is clearly seen that the data is comprised of many factor variables. Now, let's look at the data structure.

```
# data loading
library(datasets)
data(mtcars)

# data structure
print(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 ...
## NULL
```

We need some work to convert that **num** objects to their respective data type. The factor variables in this dataset are **gear**,**cyl**,**vs**,**carb** and **am**. Among those factor variables we are interested in **am** and to understand it better the level of that variable is also renamed.

```
#data wrangling
mtcars$cyl <- as.factor(mtcars$cyl); mtcars$vs <- as.factor(mtcars$vs)
mtcars$gear <- as.factor(mtcars$gear); mtcars$carb <- as.factor(mtcars$carb)
mtcars$am <- as.factor(mtcars$am)

#renaming the level
levels(mtcars$am) <- c("Auto", "Manual")
```

The relationships between the variables the scatter plots is produced to obeserve each variable against all others. That pairs graph is shown in Appendix figure 1. Those scatter plots clearly show that Mile per gallon(MPG) tends to correlate well with many of the other variables. Mile per gallon(MPG) and transmission types (am) are the particularly interesting variables. The box plot between these two variables is also shown in Appendix figure 2. This figure indicates that Manual Transmission tends to present larger values of mpg than the automatic ones.

2. Regression Analysis

Among many regression models, linear models should be chosen as the outcome we expected is neither binary nor count variable. First of all, as a direct implementation of the first question, the model is fitted with **mpg** as outcome with **am** regressor.

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

##
## 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
```

The model is statistically significant. with p value less than 0.05. however, R-sqaured value is 0.3598 which can be interpreted as '35 percent of total variation in **mpg** is explained by **am**. It is too low to keep the model. The model is fitted again with all the variables in the datasets except **mpg**.

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

When we look at the summary, Adjusted R-squared is 77.9% and it also solve the statistical significance. The first column of the Appendix figure 1 suggests that **dart**, **qesc**, **gear** and **carb** would have less impact on **mpg**. The graph also shows multicollinearity in independent variables. Correlation between numeric independent variables are as follow.

```
cor(mtcars[c('disp','hp','drat','wt','qsec')])
```

```
##            disp         hp        drat         wt        qsec
## disp  1.0000000  0.7909486 -0.71021393  0.8879799 -0.43369788
## hp    0.7909486  1.0000000 -0.44875912  0.6587479 -0.70822339
## drat -0.7102139 -0.4487591  1.00000000 -0.7124406  0.09120476
## wt    0.8879799  0.6587479 -0.71244065  1.0000000 -0.17471588
## qsec -0.4336979 -0.7082234  0.09120476 -0.1747159  1.00000000
```

Most of the values are greater then 0.7 which indicate that those are highly correlated to each other. It indicates that changes in one variable are associated with shifts in another variable. The stronger the correlation, the more difficult it is to change one variable without changing another. So, instead of refitting the model, R programming provide **step** function.This function will perform the selection by calling *lm* repeatedly. It selects the best variables to use in predicting mpg with the *Akaike information criterion* that implements both forward selection and backward elimination.

```
best_fit <- step(fit_all,direction='both',trace = 0)
summary(best_fit)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, 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 ***
## 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 **
## amManual     1.80921    1.39630   1.296  0.20646
## ---
## 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
```

Best fit eliminated most of the highly correlated variables. The models are then compared with *anova* function.

```
anova(simple_fit,best_fit,fit_all)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
## 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     26 151.03  4    569.87 17.7489 1.476e-05 ***
## 3     15 120.40 11     30.62  0.3468    0.9588
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P-value of *best_fit* model is significant compared to *fit_all* model. It means that we can confidently **reject Null Hypothesis**. Now, the coefficient of *best_fit* model is checked.

```
best_fit$coefficients
```

```
## (Intercept)        cyl6        cyl8          hp          wt    amManual
## 33.70832390 -3.03134449 -2.16367532 -0.03210943 -2.49682942  1.80921138
```

Manual transmission coefficient in this model is 1.81 which can be interpreted as the expected value of **mpg** with manual transmission is 1.81 larger than that of with automatic transmission. Confidence interval of **amManual** can also be observed.

```
confint(best_fit,'amManual')
```

```
##              2.5 %   97.5 %
## amManual -1.060934 4.679356
```

3. Residual Analysis

We now check the influence measures to the selected model.It is possible for a single observation to have a great influence on the results of a regression analysis. It is therefore important to detect influential observations and to take them into consideration when interpreting the results. *dfbetas* function gives the data points that influence the model coefficient most. We are interested data points influence on **amManual** coefficient.

```
coeff <- dfbetas(best_fit)
amManual_coef <- coeff[,6]
head(sort(amManual_coef,decreasing = TRUE))
```

```
##     Toyota Corona          Fiat 128 Chrysler Imperial    Toyota Corolla
##        0.73054020        0.42920432        0.35074579        0.28853987
```

```
##        Camaro Z28        Duster 360
##        0.08398495        0.06589986
```

These six data points are influence most but they are not greated than 1. *hatvalues* function gives the measure of leverage.

```
leverage <- hatvalues(best_fit)
head(sort(leverage,decreasing = TRUE))
```

```
##       Maserati Bora Lincoln Continental       Toyota Corona   Chrysler
Imperial
##           0.4713671           0.2936819           0.2777872
0.2611168
##       Mazda RX4 Wag  Cadillac Fleetwood
##           0.2496110           0.2496026
```

Toyota Corona and Chrysler Imperial are far from fitted line and has some impact **amManual** coefficient.

We now look at the diagnostics of our chosen model. Residual plots can be found in the Appendix figure 3. The assumptions of models are - Outcome can be expressed as linear function of regressors - Variation of observations around the regression line is constant (homoscedasticity) - Outcomes are normally distributed. In **Residual vs Fitted** graph, the red line is not much flat. It suggests that the linear assumption is not met. There is no pattern in this graph so the variation is constant. And there is no non-constant variance which means there is no **Heteroskedasticity**. For **Normal Q-Q**, the error are somewhat normally distributed. In **Scale-Location** plot, there are some potential points of interest in the plots that may indicate values of increased leverage of outliers.

4. Conclusions

In summary, we have performed fairly robust model fits although there is some potential of being non-linearity. The selected model perform and explain a lot better than simple linear model with only variable. Confiendence interval does not strongly indicate the statement of the interpretation of coefficient as described above. 95 out 100 cases, the correlation between **amManual** and **mpg** is between -1.061 and 4.8. So, it can be greater positive impact or slight negative impact on **mpg**. If we have more observations available,they could help us better answer the second question about: Quantify the MPG difference between automatica and manual transmissions? The database with only 32 observations may not have been enough to answer more clearly the second question.

APPENDIX - GRAPHICS

Figure 1 : Pairs graph

```
pairs(mtcars, panel = panel.smooth, main = "MTCARS PAIRS GRAPHS")
```

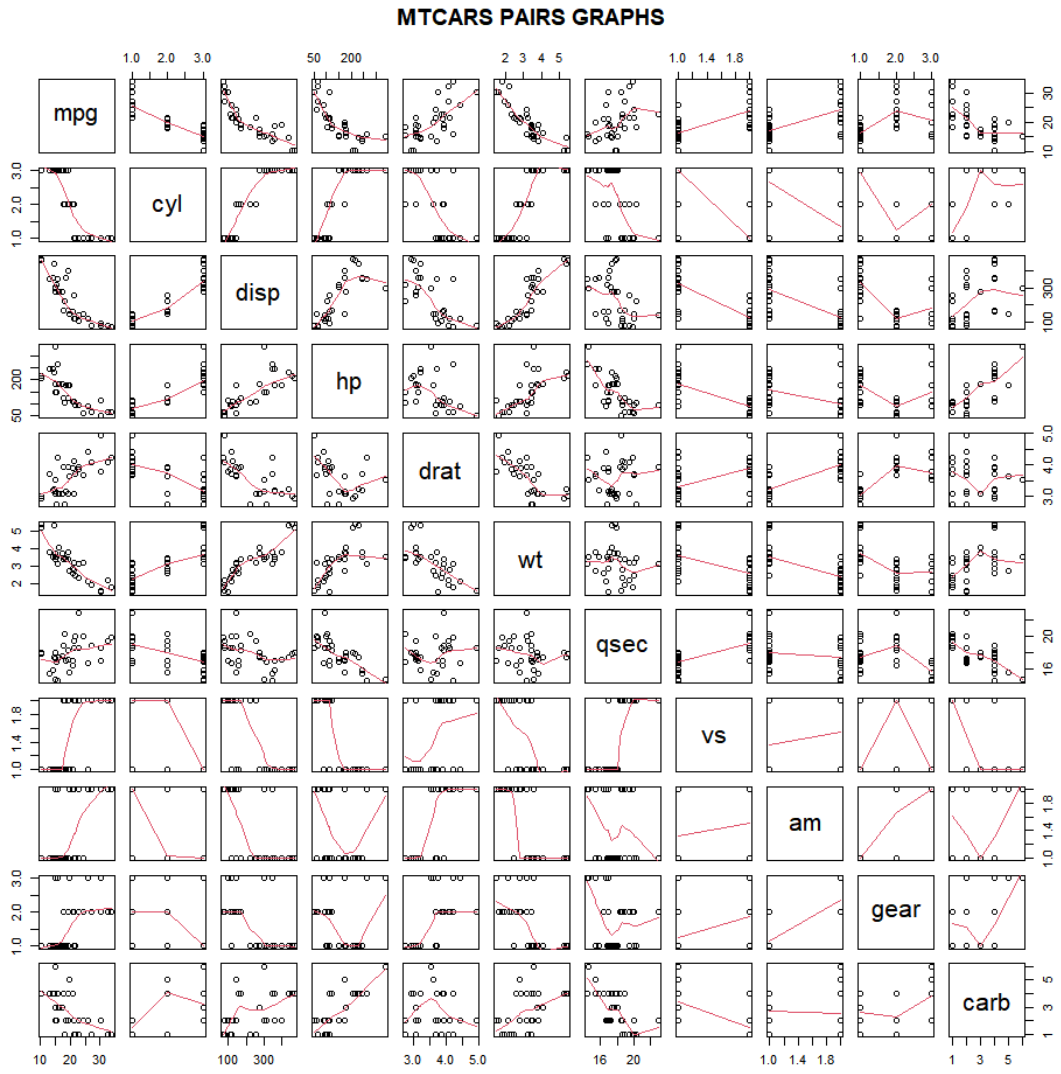


Figure 2 : Boxplots of “mpg” versus “am”

```
library(ggplot2)
transTyp <- ggplot(aes(x=am, y=mpg), data=mtcars) +
geom_boxplot(aes(fill=am))
transTyp <- transTyp + labs(title = "Automatic vs Manual Transmission
Boxplot")
transTyp <- transTyp + xlab("Transmission Type")+ ylab("MPG")
transTyp <- transTyp + labs(fill = "Transmission Types")
transTyp <- transTyp + theme(plot.title = element_text(hjust = 0.5))
transTyp
```

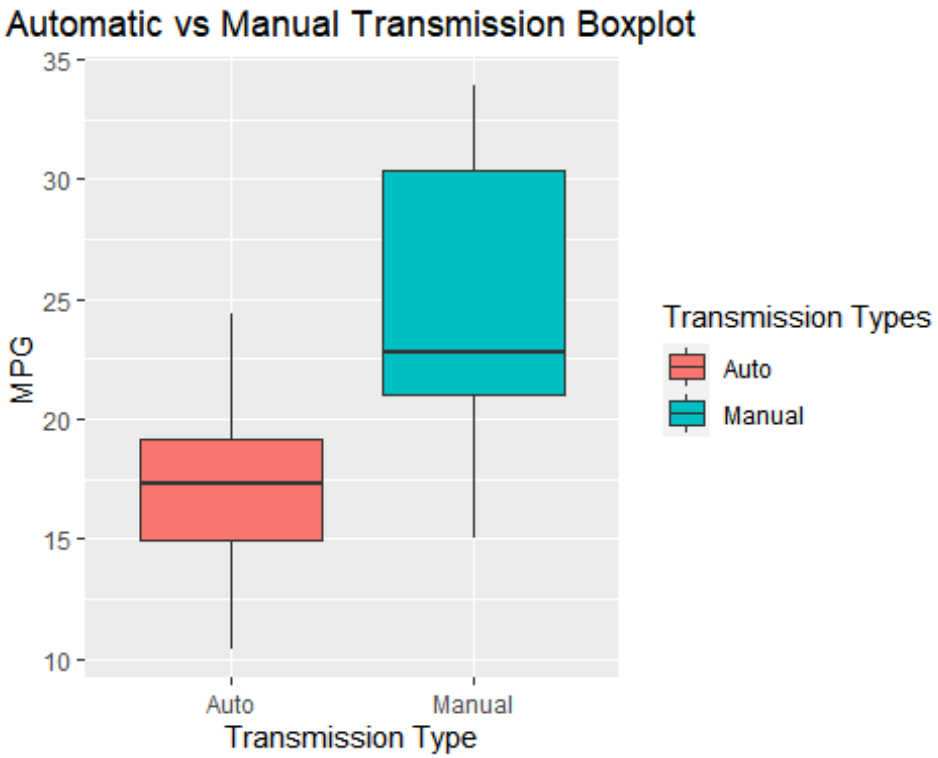
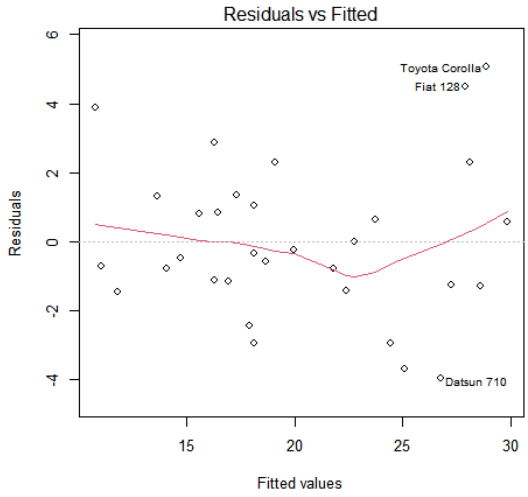
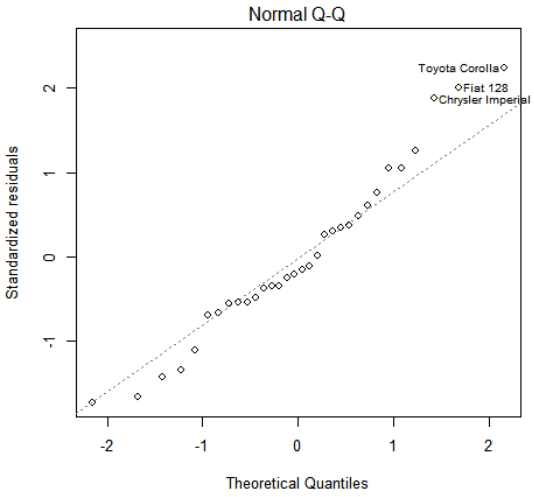
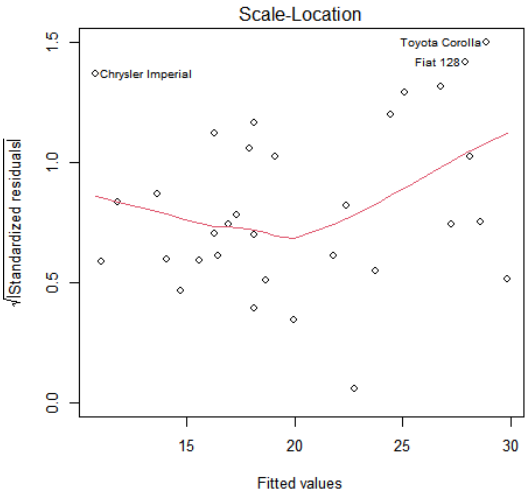


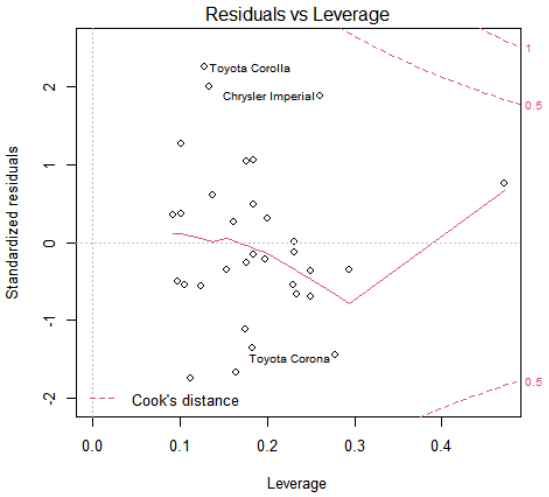
Figure 3 : Residual plots for selected model

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


Residuals vs Fitted
Toyota Corolla
Fiat 128
Datsun 710
Residuals
Fitted values
6
4
2
0
-2
-4
15
20
25
30


Normal Q-Q
Toyota Corolla
Fiat 128
Chrysler Imperial
Standardized residuals
Theoretical Quantiles
2
1
0
-1
-2


Scale-Location
Toyota Corolla
Fiat 128
Chrysler Imperial
√|Standardized residuals|
Fitted values
1.5
1.0
0.5
0.0
15
20
25
30


Residuals vs Leverage
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
Standardized residuals
Leverage
1
0.5
0.5
2
1
0
-1
-2
0.0
0.1
0.2
0.3
0.4