Regression Models Project Report - Motor Trend Data Analysis

Ravi Kumar Yadav

Executive Summary

In this project detailed Analysis of the mtcars data has been performed. Main objective of the research is to quntify:

1. Is an automatic or manual transmission better for MPG?
2. Quantifying how different is the MPG between automatic and manual transmissions?

Regression models and exploratory data analyses is uesed to mainly explore how automatic and manual transmissions features affect the MPG feature.

Exploratory Data Analysis

In this secsion some exploratory data analysis is performed for the mtcars data. First we load the data, display the head, did data transformation for necessary variables from numerics to factors . Attach function is used to attach the database to R search path, so objects in the database can be accessed by simply giving their names.. Then Structure of the transformed data is available in appendix #1.

```r
library(ggplot2)
data(mtcars)
head(mtcars) # Sample Data
dim(mtcars)
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$am <- factor(mtcars$am,labels=c('Automatic','Manual'))
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
attach(mtcars)
```

```
## The following object is masked from package:ggplot2:
##
##     mpg
```

Plots fpr the exploratory data analysis is added in the **Appendix**. Box plot very clearly indicates the higher level of MPG for Manual transmission type. And in the Pair plot we can see higher correlation between variable like "wt", "hp", "disp".

Inference

In this section t.test is performed for the NULL hypothesis as MPG for Automatic and Manual Transsmission are from same same population assuming MEG has a normal distribution.

```
result <- t.test(mpg ~ am)
result$p.value

## [1] 0.001373638

result$conf

## [1] -11.280194  -3.209684
## attr(,"conf.level")
## [1] 0.95

result$estimate

## mean in group Automatic    mean in group Manual
##                17.14737                24.39231
```

Confidence interval does not include zero and the p value is 0.001373638, so the NUll hypothesis an be easily rejected. Mean for MPG value is almost 7 more for manual transmission type than automatic transmission.

Regression Analysis

In this section, we start building linear regression models based on the different variables and try to find out the best model fit and compare it with the base model which we have using anova

based on the pairs plot where several variables seem to have high correlation with mpg, We build an initial model with all the variables as predictors, and perfom stepwise model selection to select significant predictors for the final model which is the best model. This is taken care by the step method which runs lm multiple times to build multiple regression models and select the best variables from them using both forward selection and backward elimination methods.

```
init_model <- lm(mpg ~ ., data = mtcars)
best_model <- step(init_model, direction = "both")
```

Please check the summary of the best model in the appendix #4. From the best model details, we observe that the adjusted R2 value is 0.84 which is the maximum obtained considering all combinations of variables. Thus, we can conclude that more than 84% of the variability is explained by the above model.

Next, we fit the simple model with MPG as the outcome variable and Transmission as the predictor variable

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

Please check appendix #5 for summary.The Adjusted R-squared value is 0.3385, which means that the model can explain about 34% of the variance of the MPG variable. The low Adjusted R-squared value also indicates that we need to add other variables to the model.

Please refer to the Appendix #6. According to the scatter plot, it indicates that there appear to be an interaction term between "wt" variable and "am" variable, since automatic cars tend to weigh heavier than manual cars. Thus, we have the following model including the interaction term:

```
amIntWtModel<-lm(mpg ~ wt + qsec + am + wt:am, data=mtcars)
```

Check the summary of this model in appendix # 7. Adjusted R-squared value is 0.8804, which means that the model can explain about 88% of the variance of the MPG variable. This is the a very good model, so we end up selecting this as the best model "mpg ~ wt + qsec + am + wt:am".

Residuals and Diagnostics

Please refer to the appendix #8 for the plots. According to the residual plots, we can verify the following underlying assumptions:
1. The Residuals vs. Fitted plot shows no consistent pattern, supporting the accuracy of the independence assumption.
2. The Normal Q-Q plot indicates that the residuals are normally distributed because the points lie closely to the line.
3. The Scale-Location plot confirms the constant variance assumption, as the points are randomly distributed.
4. The Residuals vs. Leverage argues that no outliers are present, as all values fall well within the 0.5 bands.

Therefore, the above analyses meet all basic assumptions of linear regression and asnwers the question as well.

Appendix

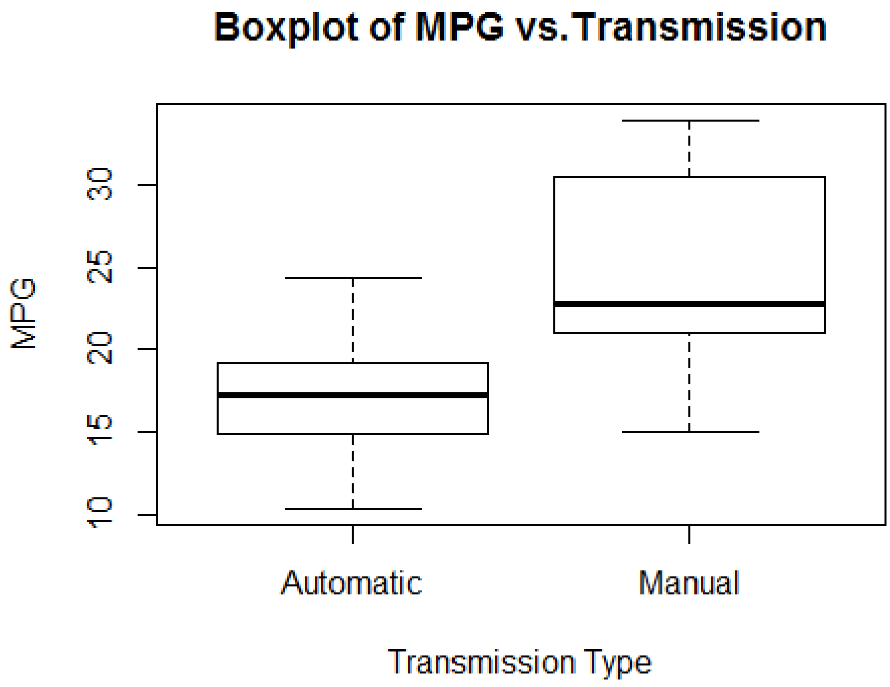
1. Structure of the transformed data

```
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 : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ 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  : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am  : Factor w/ 2 levels "Automatic","Manual": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ carb: Factor w/ 6 levels "1","2","3","4",..: 4 4 1 1 2 1 4 2 2 4 ...
```

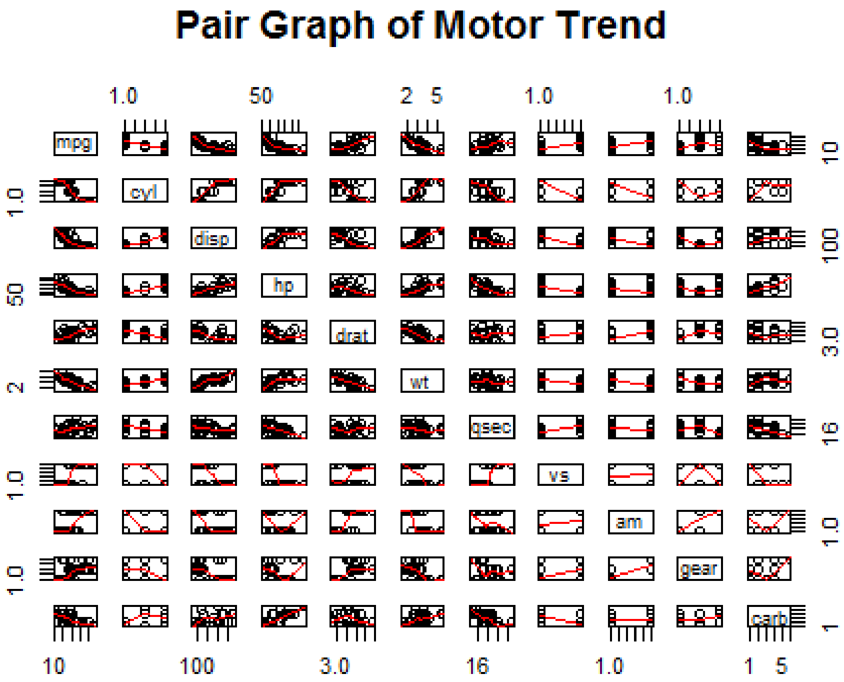
2. Boxplot of MPG vs. Transmission Type

```
boxplot(mpg ~ am, xlab="Transmission Type", ylab="MPG", main="Boxplot of MPG
vs.Transmission")
```



3. Pair Graph of The Motor Trend

```
pairs(mtcars, panel=panel.smooth, main="Pair Graph of Motor Trend")
```



4.Summay of the best regression model

```
summary(best_model)

##
## 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
```

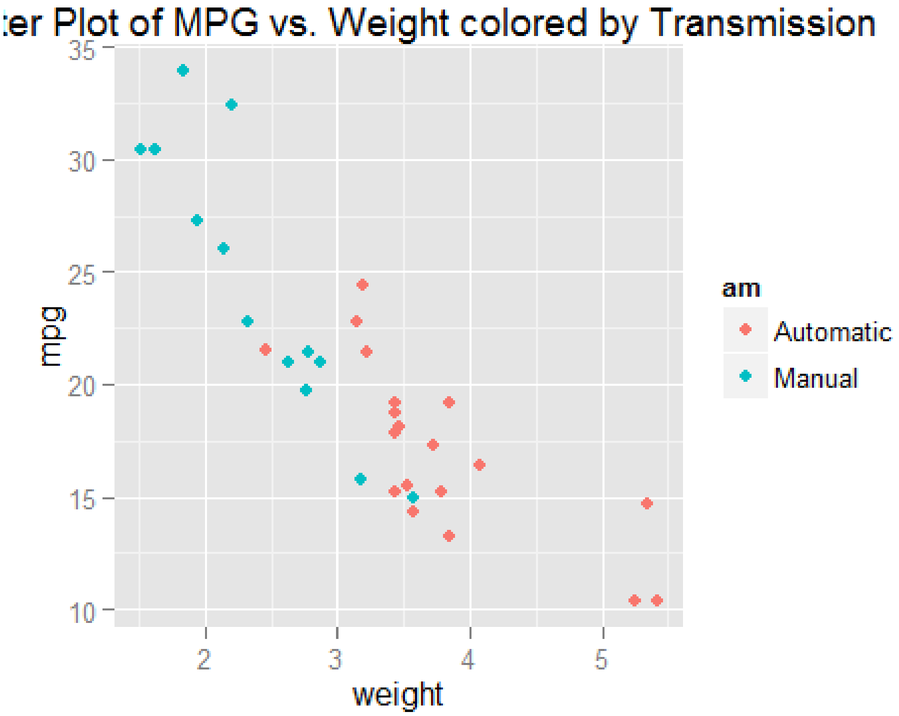
5.Summay of the regression model with am as predictor

```
summary(amModel)

##
## 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
```

6.Scatter Plot of MPG vs. Weight colored by Transmission

```
ggplot(mtcars, aes(x=wt, y=mpg, group=am, color=am, height=3, width=3)) +
geom_point() +
scale_colour_discrete(labels=c("Automatic", "Manual")) +
xlab("weight") + ggtitle("Scatter Plot of MPG vs. Weight colored by
Transmission")
```



7.Summay of the regression model with am as predictor

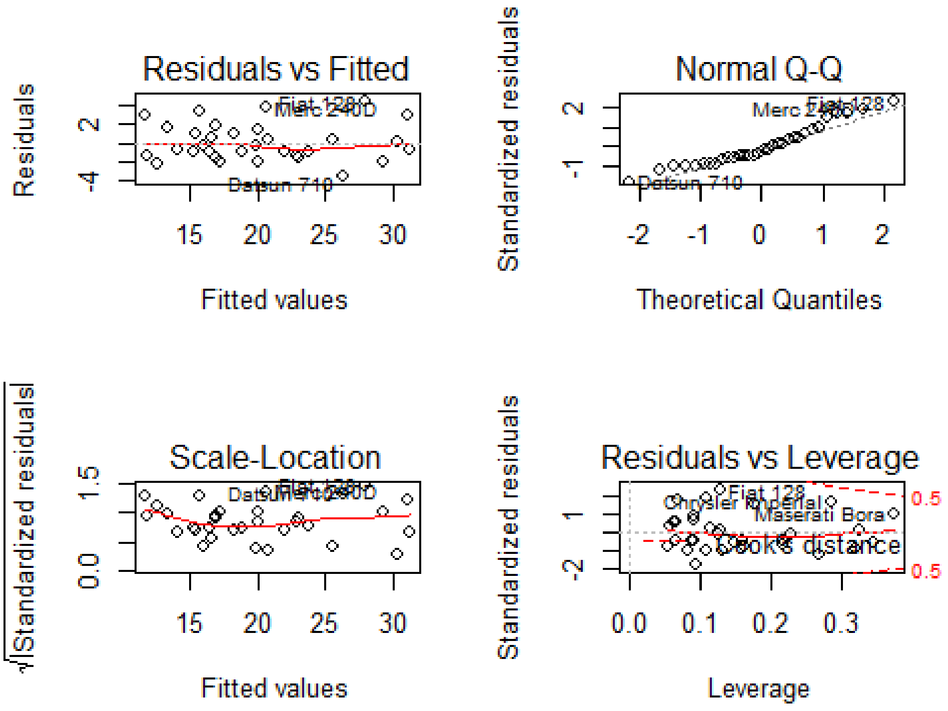
```
summary(amIntWtModel)

##
## Call:
## lm(formula = mpg ~ wt + qsec + am + wt:am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5076 -1.3801 -0.5588  1.0630  4.3684
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.723      5.899   1.648 0.110893
## wt            -2.937      0.666  -4.409 0.000149 ***
## qsec           1.017      0.252   4.035 0.000403 ***
## amManual      14.079      3.435   4.099 0.000341 ***
## wt:amManual   -4.141      1.197  -3.460 0.001809 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## 
## Residual standard error: 2.084 on 27 degrees of freedom
## Multiple R-squared:  0.8959, Adjusted R-squared:  0.8804 
## F-statistic: 58.06 on 4 and 27 DF,  p-value: 7.168e-13
```

8.Residual Plots

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


Residuals vs Fitted
Residuals
2
-4
Fiat 128
Merc 240D
Datsun 710
15
20
25
30
Fitted values
Normal Q-Q
Standardized residuals
2
-1
Merc 240D
Fiat 128
Datsun 710
-2
-1
0
1
2
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
1.5
0.0
Datsun 710
Fiat 128
Merc 240D
15
20
25
30
Fitted values
Residuals vs Leverage
Standardized residuals
1
-2
Fiat 128
Chrysler Imperial
Maserati Bora
Cook's distance
0.5
0.5
0.0
0.1
0.2
0.3
Leverage