Regression Models For Motor Trends

Kumar Shaket

01/03/2021

Executive Summary

In this report, we will examine the mtcars data set and explore how miles per gallon (MPG) is affected by different variables. In particularly, we will answer the following two questions:

(1) *Is an automatic or manual transmission better for MPG*
(2) *Quantify the MPG difference between automatic and manual transmissions*

Exploratory Analysis

Loading required libraries and dataset=mtcars

```
library(ggplot2)
library(datasets)
data(mtcars)
head(mtcars,3)
```

```
##                mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4     21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag 21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710    22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
```

Transforming certain variables to factor variables

```
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
mtcars$am <- factor(mtcars$am, labels = c("Automatic","Manual"))
head(mtcars,5)
```

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

Regression Analysis

We see that manual transmission is better for mpg than automatic transmission for cars as seen below.

```
aggregate(mpg ~ am, data=mtcars,mean)
```

```
##          am      mpg
## 1 Automatic 17.14737
## 2    Manual 24.39231
```

We hypothesize that automated transmission has approx 7 mpg lower than manual transmission. To determine if there is significant difference, let's do t.test.

```
t_automatic <- mtcars[mtcars$am=="Automatic",]
t_manual <- mtcars[mtcars$am=="Manual",]
t.test(t_automatic$mpg,t_manual$mpg)
```

```
##
##  Welch Two Sample t-test
##
## data:  t_automatic$mpg and t_manual$mpg
## 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 of x mean of y
##  17.14737  24.39231
```

The p-value is 0.001374, thus we can state this is a significant difference. Now to quantify this, we will build

Linear Models

Model1

```
mt1 <- lm(mpg ~ am,mtcars)
summary(mt1)
```

```
##
## 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
```

This shows us that the average MPG for automatic is 17.1 MPG, while manual is 7.2 MPG higher via boxplot(Appendix - Plot 1)

The R2 value is 0.36 thus telling us model only explains us 36% of the variance. As a result, we need to build a multivariate linear regression.

Model2

The new model will use the other variables to make it more accurate. We explore the other variable via a pairs plot (Appendix - Plot 2) to see how all the variables correlate with mpg.

```
mt2 <- lm(mpg ~ am + cyl + disp + hp + wt,mtcars)
summary(mt2)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + disp + hp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9374 -1.3347 -0.3903  1.1910  5.0757
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.864276   2.695416  12.564 2.67e-12 ***
## amManual     1.806099   1.421079   1.271   0.2155
## cyl6        -3.136067   1.469090  -2.135   0.0428 *
## cyl8        -2.717781   2.898149  -0.938   0.3573
## disp         0.004088   0.012767   0.320   0.7515
## hp          -0.032480   0.013983  -2.323   0.0286 *
## wt          -2.738695   1.175978  -2.329   0.0282 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.453 on 25 degrees of freedom
## Multiple R-squared:  0.8664, Adjusted R-squared:  0.8344
## F-statistic: 27.03 on 6 and 25 DF,  p-value: 8.861e-10
```

From this we see that cyl, disp, hp, wt have the strongest correlation with mpg.

We will compare both the models using anova function.

```
anova(mt1,mt2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl + disp + hp + wt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     25 150.41  5    570.49 18.965 8.637e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This results in a p-value of 8.637e-08, and we can claim the mt2 model is significantly better than our mt1 linear model.

```r
summary(mt2)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + disp + hp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9374 -1.3347 -0.3903  1.1910  5.0757
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.864276   2.695416  12.564 2.67e-12 ***
## amManual     1.806099   1.421079   1.271   0.2155
## cyl6        -3.136067   1.469090  -2.135   0.0428 *
## cyl8        -2.717781   2.898149  -0.938   0.3573
## disp         0.004088   0.012767   0.320   0.7515
## hp          -0.032480   0.013983  -2.323   0.0286 *
## wt          -2.738695   1.175978  -2.329   0.0282 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.453 on 25 degrees of freedom
## Multiple R-squared:  0.8664, Adjusted R-squared:  0.8344
## F-statistic: 27.03 on 6 and 25 DF,  p-value: 8.861e-10
```

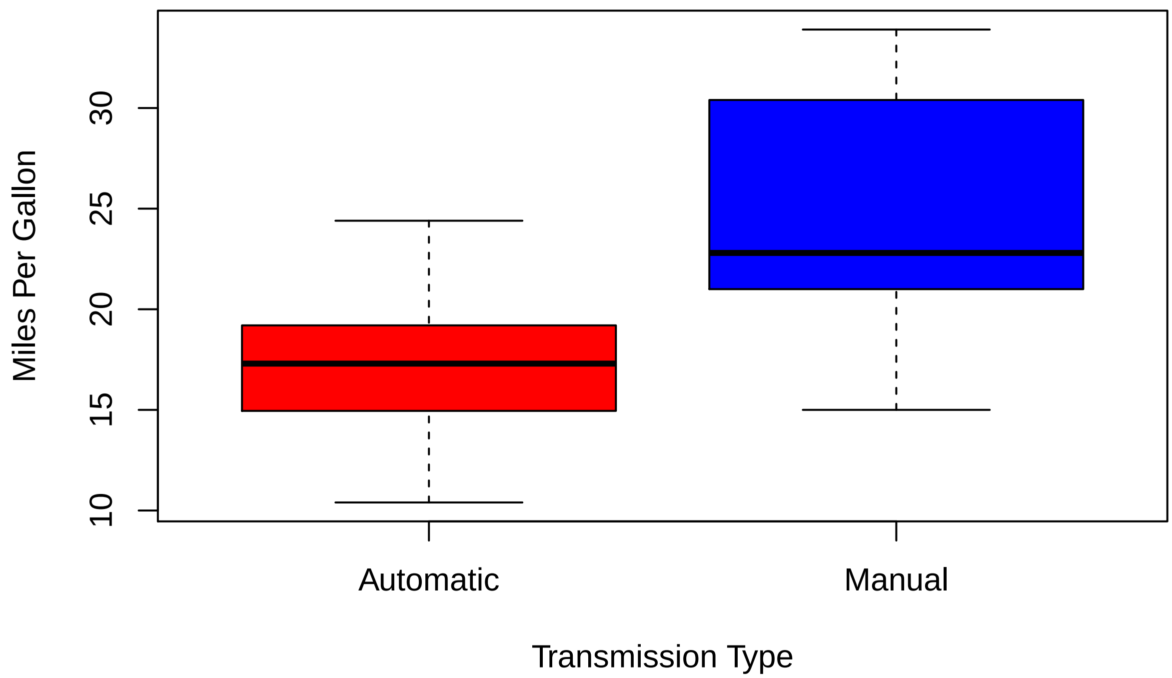
The model explains 86.64% of the variance and as a result, cyl, disp, hp, wt did affect the correlation between mpg and am.

Thus, we can say the difference between automatic and manual transmissions is 1.81 MPG.

Appendix

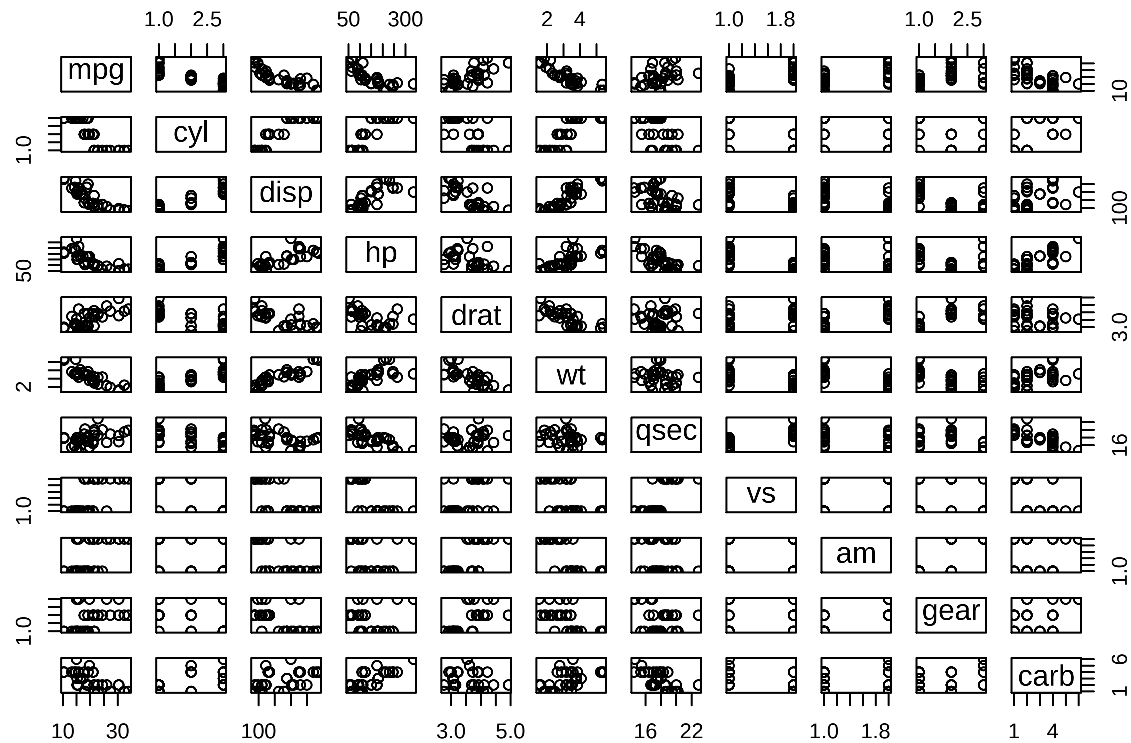
Plot 1 : Boxplot of MPG by transmission type

```r
boxplot(mpg ~ am,data = mtcars, col = (c("red","blue")), ylab = "Miles Per Gallon", xlab = "Transmissio
```



Plot 2 - Pairs plot for the data set

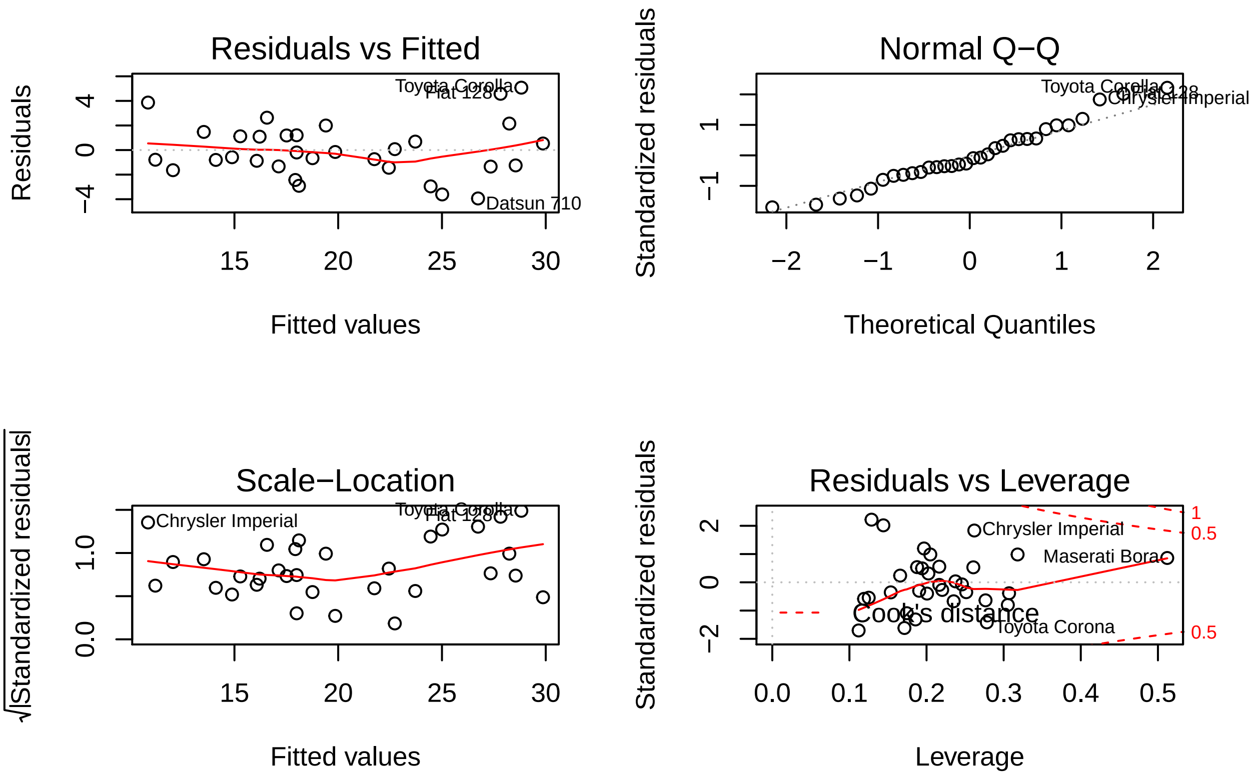
```r
pairs(mpg ~ ., data = mtcars)
```



We double-check the residuals for non-normality (Appendix - Plot 3) and can see they are all normally distributed.

Plot 3 - Check residuals

```
par(mfrow = c(2,2))
plot(mt2)
```


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Datsun 710
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Fiat 128
Chrysler Imperial
Scale−Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals vs Leverage
Standardized residuals
Leverage
Chrysler Imperial
Maserati Bora
Toyota Corona
Cook's distance
1
0.5
0.5