Regression Models Course Project

Anjali Singh

8 October 2017

Executive Summary

In this project we will evaulate the mtcars data set and explore how miles per gallon (MPG) is affected by different variables. We will aim to answer the below two question

1. Is an automatic or manual transmission better for MPG?

2. Quantify the MPG difference between automatic and manual transmissions.

Conclusion:

Manual Transimission is much better than in Automatic Transmission in giving good mileage

1. Car with automatic transmissions gives 17.147 MPG, and with manual transmissions car gives 24.392 MPG.

2. In the multiple regression model, the MPG difference is 2.9358 MPG at the mean weight and qsec.

Exploratory Analysis

```
library(ggplot2)
data(mtcars)
summary(mtcars$mpg[mtcars$am==0])
str(mtcars)
```

Converting into factors

```
mtcars$am <- as.factor(mtcars$am)
```

There are 32 observation, 10 control variable and 1 dependent varaible in this dataset, we will first try and do some exploratory data analysis to pick up any trend. We will first check for the mean for both Auto and Manual transimission, 1 is Manual and 0 is Auto. As you see below the Mean for auto is less than that of manual. This is our first indication that Manual Transimission is better

```
sapply((split(mtcars$mpg, mtcars$am)), mean)
```

```
##        0        1 
## 17.14737 24.39231
```

Now lets try and understand this by plotting the box plot(Apendix 1) The box plot above clearly tells us that Manual Transimission provides much better transmission than Automatic Transimission. Lets do a t test to hypothesis to see if there is impact of Transimission type on mileage

Hypothesis:

H0:Transimission Type has no impact on Mileage

H1:Transimission Type has a impact on Mileage

```
t.test(mtcars[mtcars$am == "0",]$mpg, mtcars[mtcars$am == "1",]$mpg)
```

As you can see, the p- Value (0.001374) is less than 0.05, therefore we reject the null hypothesis and can safely say that Transimission Type has a impact on Mileage.

Regression Analysis

Let us first try and do a linear model to quantify the relationship between Transimission and Mileage.

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

THe above model only account for 36 percent of the variation in the data, which is not a good fit. This can happen if there are more variable to contributes to the variation of the model.Lets try and capture that using the Mutliple regression model

Mulitple Regression

We will first try and fit the full model and try and get the predictors that are important for us to capture

```
initial <- lm(mpg ~ ., data = mtcars)
best <- step(initial, direction = "both")
summary(best)
```

```
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 ***
## am1           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
```

The three variables for the model are qsec, wt, and am, The R squared value was 84 % and Residual standard error: 2.459 on 28 degrees of freedom. P value for beta3(am) is 0.047 and is small enough for us to reject the null hypothesis and safely say that Transmission does impact mileage.

Residuals and Diagnostics(Appendix 2)

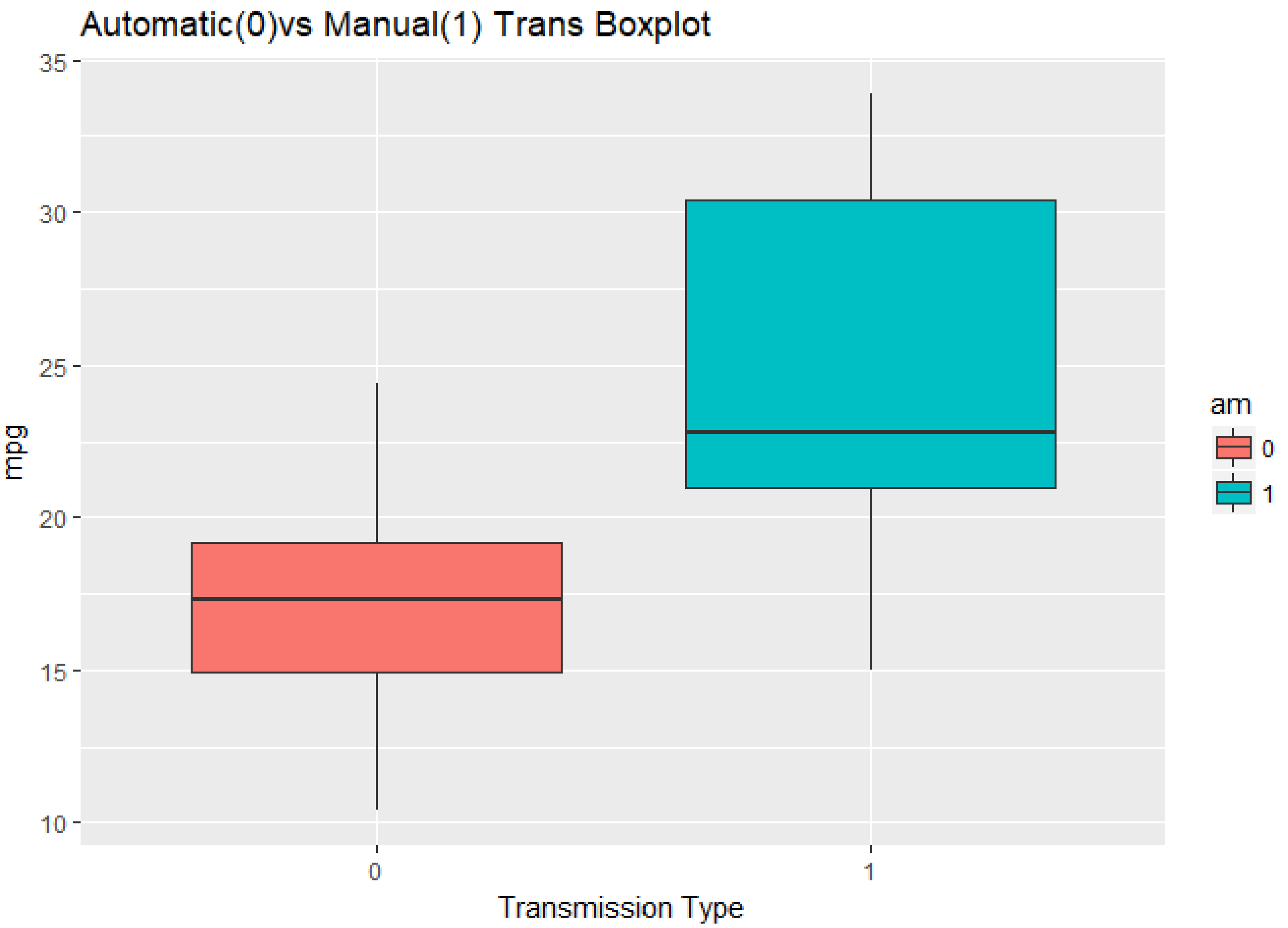
The Normal Q-Q plot indicates the normal distribution of the date while residual vs fitted shows the data us random and and is independent. Scale location plot indicates spread of points across predicted values range and here the residuals have uniform variance across the range.There is no Homoscedasticity The residuals for the Chrysler Imperial, Fiat 128, and Toyota Corolla are outliers and exert some influence on the curve

Apendices

Apendix 1

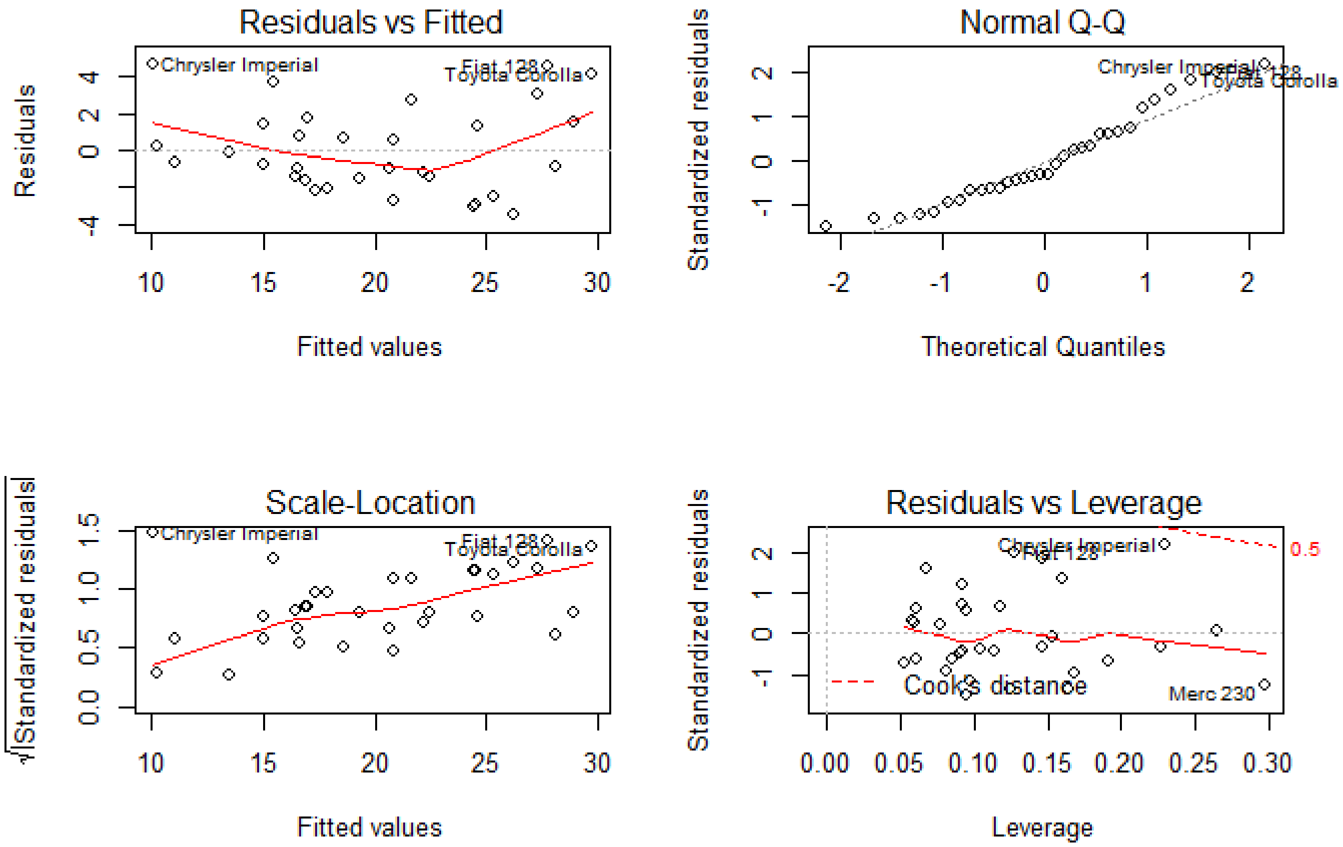
```
#Visualization ~ Automatic vs Manual Transmission:
library(ggplot2)

g <- ggplot(aes(x=am, y=mpg), data=mtcars) + geom_boxplot(aes(fill=am))
g <- g + labs(title = "Automatic(0)vs Manual(1) Trans Boxplot")
g <- g + xlab("Transmission Type")
g
```



Apendix 2

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


Residuals vs Fitted
Residuals
Fitted values
-4
0 2
4
10
15
20
25
30
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q-Q
Standardized residuals
Theoretical Quantiles
-1
0
1
2
-2
-1
0
1
2
Chrysler Imperial
Fiat 128
Toyota Corolla
Scale-Location
√|Standardized residuals|
Fitted values
0.0
0.5
1.0
1.5
10
15
20
25
30
Chrysler Imperial
Fiat 128
Toyota Corolla
Residuals vs Leverage
Standardized residuals
Leverage
-1
0
1
2
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Chrysler Imperial
Fiat 128
Merc 230
Cook's distance
0.5