Regression Model

Sandesh

10/28/2019

Management Summary

Motor Trend, an automobile trend magazine is interested in exploring the relationship between a set of variables and miles per gallon (MPG) outcome. In this project, we will analyze the mtcars dataset from the 1974 Motor Trend US magazine to answer the following questions:

```
Is an automatic or manual transmission better for miles per gallon (MPG)?
How different is the MPG between automatic and manual transmissions?
```

Using simple linear regression analysis, we determine that there is a signficant difference between the mean MPG for automatic and manual transmission cars. Manual transmissions achieve a higher value of MPG compared to automatic transmission. This increase is approximately 2.1 MPG when switching from an automatic transmission to a manual one, with the weight, horsepower and discplacement held constant.

Exploratory analysis and visualizations are located in the Appendix to this document.

Exploratory data analysis

Lets have a look at our dataset

```r
library(ggplot2) #for plots
```

```
## Warning: package 'ggplot2' was built under R version 3.6.1
```

```r
data(mtcars) #loading the dataset
head(mtcars) #viewing first few rows of the dataset
```

```
##                    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
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```r
# Transform certain variables into factors
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"))
```

To help us understand the data, we build exploratory plots. Appendix - Plot 1, shows there is a definite impact on MPG by transmission with Automatic transmissions having a lower MPG.

Regression Analysis

We've visually seen that automatic is better for MPG, but we will now quantify his difference.

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

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

Thus we hypothesize that automatic cars have an MPG 7.25 lower than manual cars. To determine if this is a significant difference, we use a t-test.

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

```
##
##  Welch Two Sample t-test
##
## data:  D_automatic$mpg and D_manual$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.

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

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

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. From this we see that cyl, disp, hp, wt have the strongest correlation with mpg. We build a new model using these variables and compare them to the initial model with the anova function.

```r
betterFit <- lm(mpg~am + cyl + disp + hp + wt, data = mtcars)
anova(init, betterFit)
```

```
## 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 betterFit model is significantly better than our init simple model. We double-check the residuals for non-normality (Appendix - Plot 3) and can see they are all normally distributed and homoskedastic.

```r
summary(betterFit)
```

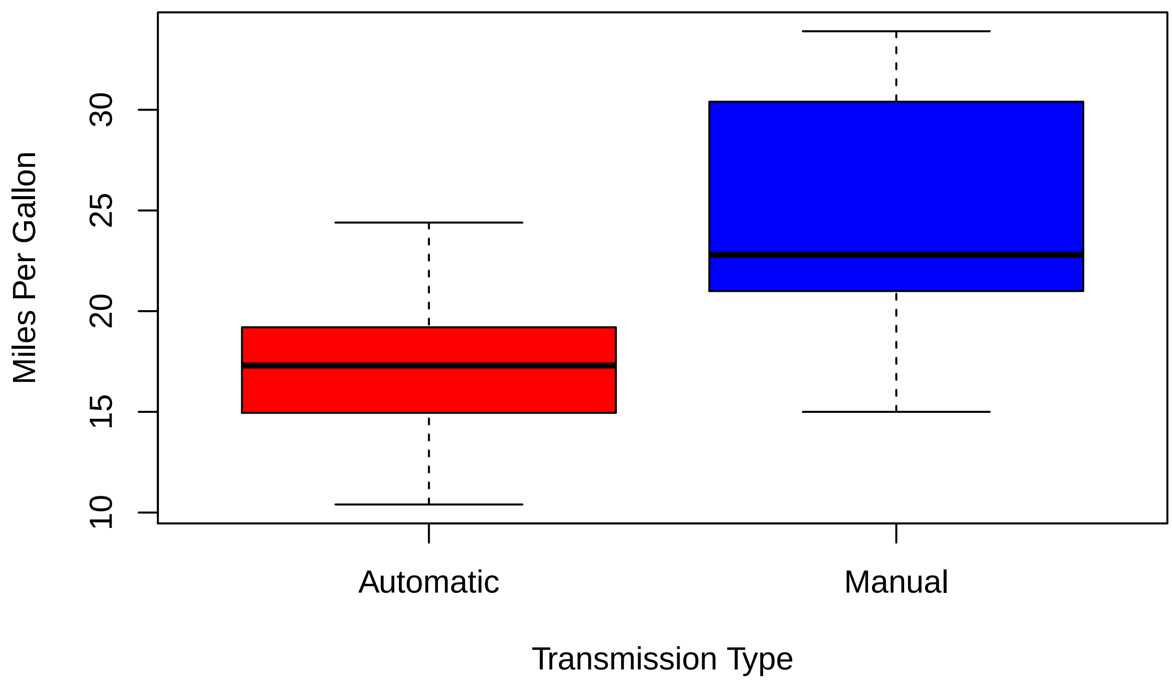
```
##
## Call:
## lm(formula = mpg ~ am + cyl + disp + hp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9374 -1.3347 -0.3903  1.1910  5.0757
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.864276   2.695416  12.564 2.67e-12 ***
## 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

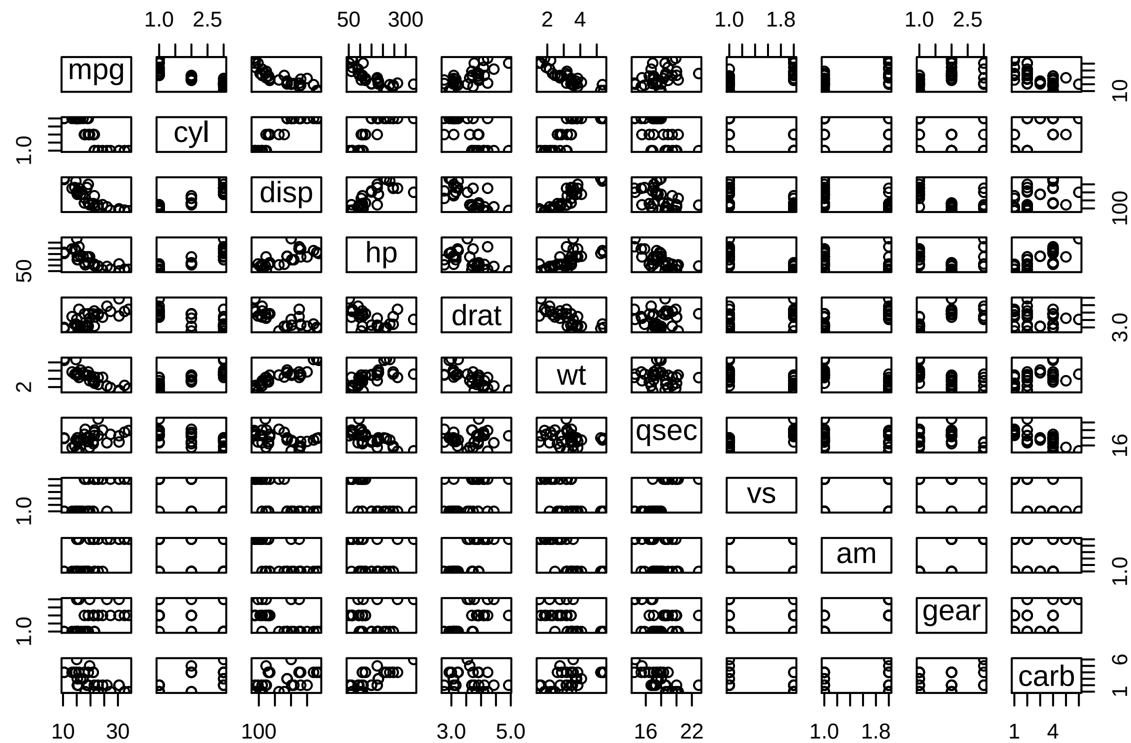
Plot 1 - Boxplot of MPG by transmission type

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



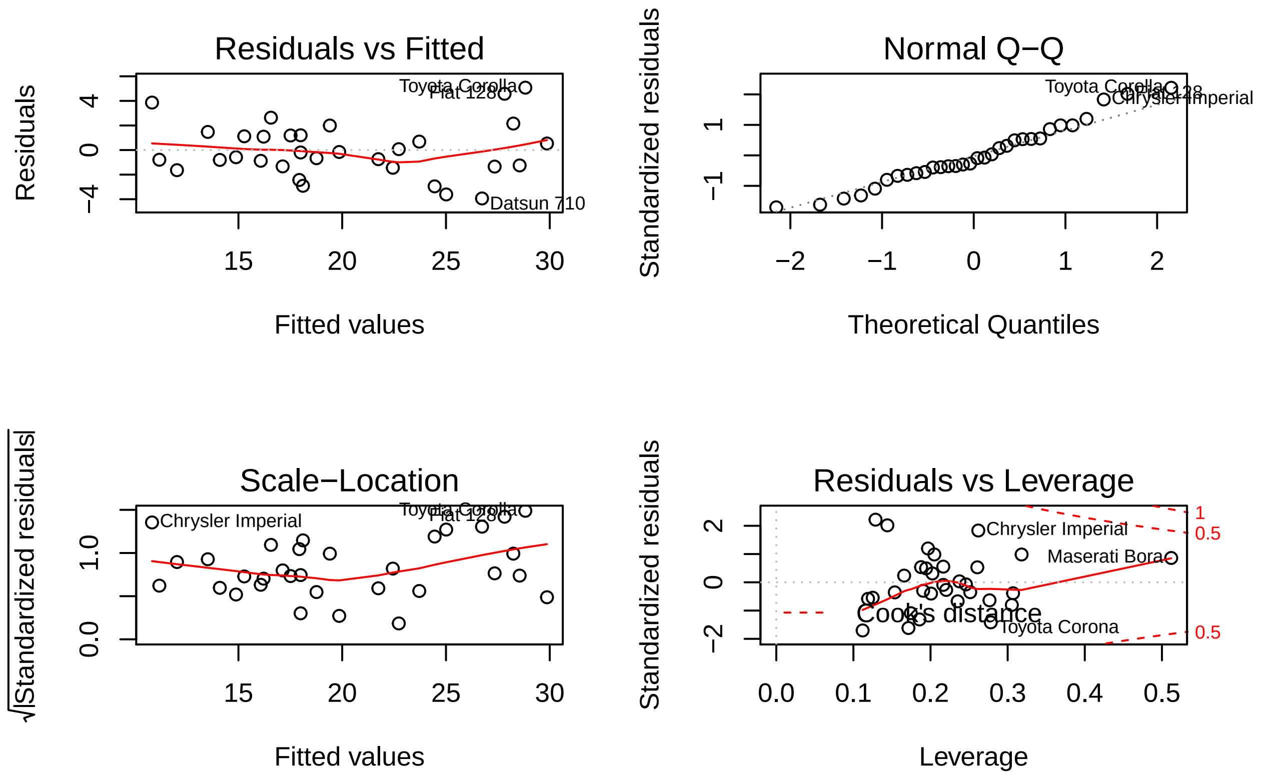
Plot 2 - Pairs plot for the data set

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



Plot 3 - Check residuals

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


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