Regression Models Course Project

Andrew Mendonca

July 28, 2017

Executive Summary

Motor Trend is a magazine about the automobile industry. They are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They want to know if an automatic or manual is transmission better for MPG and the MPG difference between automatic and manual transmissions. In this report, we will evaluate the mtcars data set and develop an analysis to answer these questions based on regression models and exploratory data analyses.

Exploratory Data Analysis

Load the following library and data set.

```r
library(ggplot2)
data(mtcars)
```

View a sample of the data set.

```r
head(mtcars)
```

```
##                    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
```

Convert variables into factors.

```r
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$am <- factor(mtcars$am, labels = cbind("Automatic", "Manual"))
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
```

We create exploratory plots to help us understand the data better. Appendix - Plot 1, illustrates the automatic transmissions having a low MPG than the manual transmissions.

Regression Analysis

View the difference between automatic and manual transmissions.

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

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

We create a hypothesis that automatic cars have a 7.25 MPG lower than manual cars. We then use a t-test.

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

```
##
##  Welch Two Sample t-test
##
## data:  mtcars[mtcars$am == "Automatic", ]$mpg and mtcars[mtcars$am == "Manual", ]$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
```

It is shown that the p-value is 0.001374, which is a significant difference. Let's quantify this.

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

```
##
## 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 average MPG for an automatic transmission is 17.1 MPG, while the manual transmission is 7.2 MPG higher. The R-squared value is 0.36, which means this model only explains 36% of the variance, so we need a multivariable model.

Appendix - Plot 2, shows how all the other variables correlate with `mpg`. We see that `cyl`, `disp`, `hp`, and `wt` have strongest correlations with mpg. We include these variables in the new model and compare them with the initial model.

```
anova(lm(mpg~am, data = mtcars), lm(mpg~am + cyl + disp + hp + wt, data = mtcars))
```

```
## 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
```

The p-value is 8.637e-08, so this multivariable model is significantly better than the linear model. Appendix - Plot 3, checks the residuals for non-normality and shows that they are all normally distributed.

```
summary(lm(mpg~am + cyl + disp + hp + wt, data = mtcars))
```

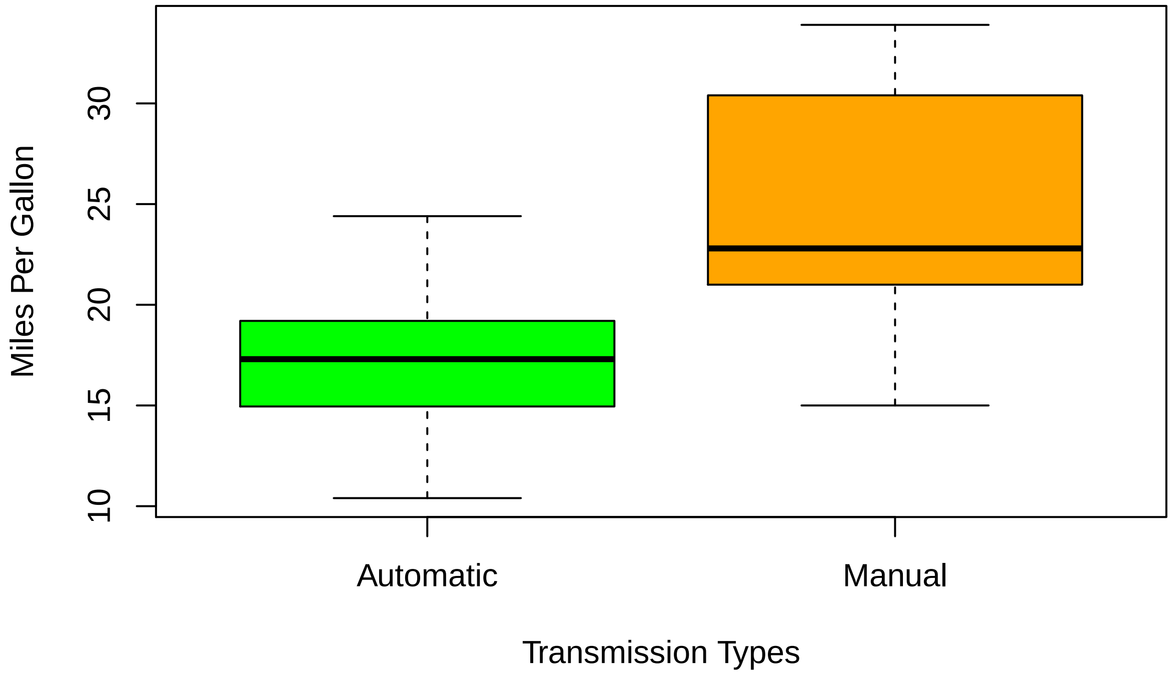
```
##
## Call:
## lm(formula = mpg ~ am + cyl + disp + hp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9374 -1.3347 -0.3903  1.1910  5.0757
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.864276   2.695416  12.564 2.67e-12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

This model only explains 86.64% of the variance, so this means that `cyl`, `disp`, `hp`, and `wt` had an effect on the correlation between `mpg` and `am`. Hence, the difference between automatic transmissions and manual transmissions is 1.81 MPG.

Appendix

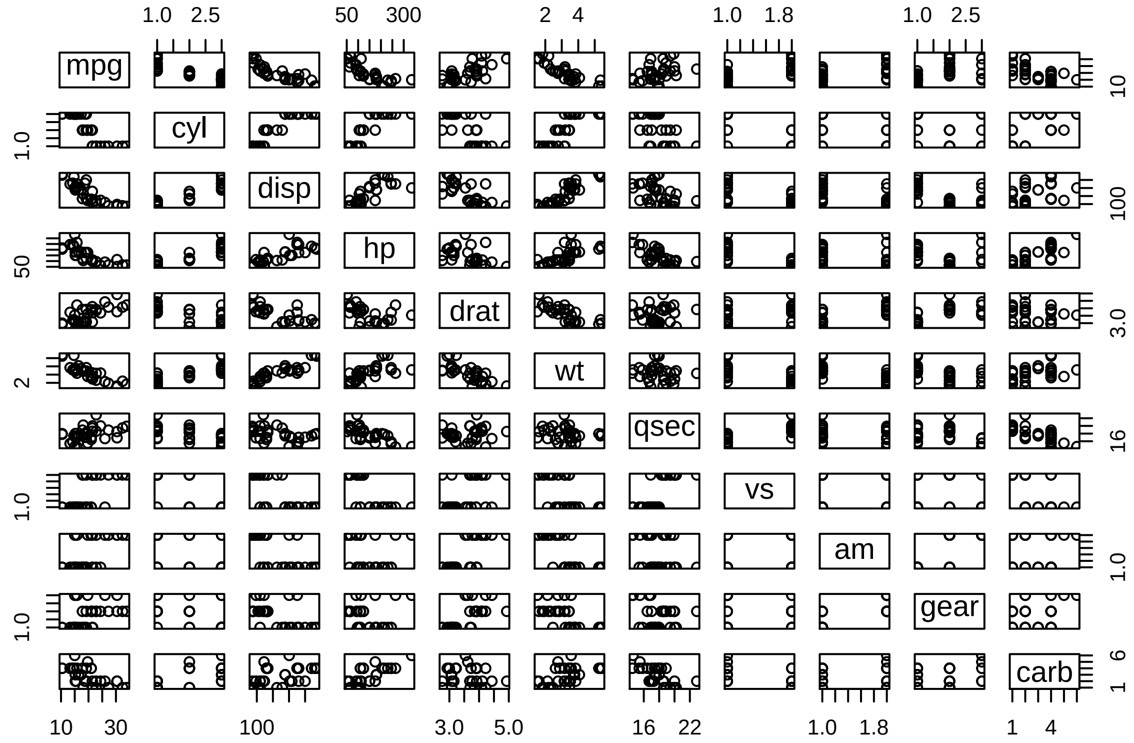
Plot 1 - Boxplot of MPG based on transmission type

```
boxplot(mpg ~ am, data = mtcars, col = (c("green", "orange")), xlab = "Transmission Types",
        ylab = "Miles Per Gallon")
```



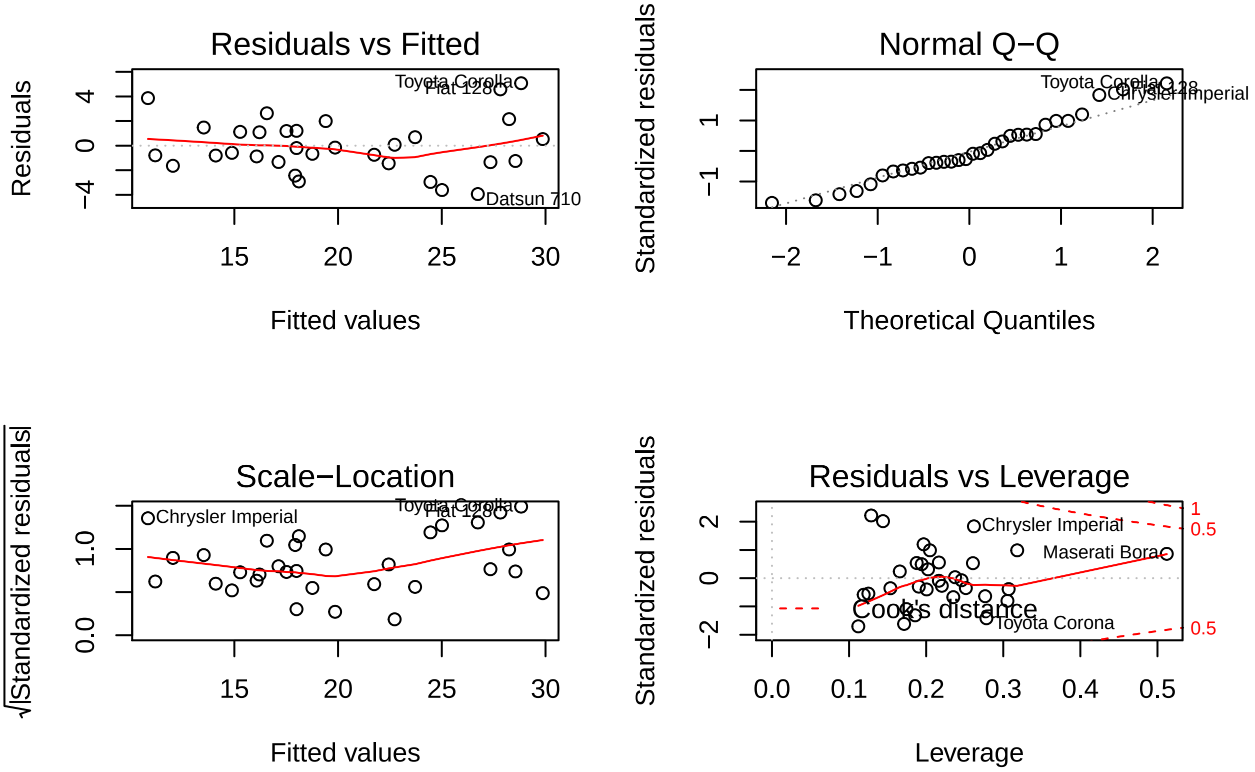
Plot 2 - Pairs plot for the rest of the data set

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



Plot 3 - Checking each residual

```
par(mfrow = cbind(2,2))
plot(lm(mpg~am + cyl + disp + hp + wt, data = mtcars))
```


Residuals vs Fitted
Residuals
Fitted values
−4
0
4
15
20
25
30
Toyota Corolla
Fiat 128
Datsun 710
Normal Q−Q
Standardized residuals
Theoretical Quantiles
−1
1
−2
−1
0
1
2
Toyota Corolla
Fiat 128
Chrysler Imperial
Scale−Location
√|Standardized residuals|
Fitted values
0.0
1.0
15
20
25
30
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals vs Leverage
Standardized residuals
Leverage
−2
0
2
0.0
0.1
0.2
0.3
0.4
0.5
Chrysler Imperial
Maserati Bora
Toyota Corona
Cook's distance
1
0.5
0.5