Regression Models Assignment - Motor Trend

R3M79

29 de Janeiro de 2018

Synopsis

In this project we assume that we're performing an analysis for the magazine Motor Trend, where we'll explore the relation between MPG (miles per galon) and different predictors

Motor Trend Analysis

Overview

Motor Trend intends to analyze the relation between MPG (miles per galon) and different variables. The goal is to answer the following questions:

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

Exploratory Analysis

First we'll load the required libraries and data (mtcars)

Data Detail and Summary

Data can be visualized in the boxplot on Appendix - Figures

```
#Data Summary
head(mydata)
```

```
##    mpg cyl disp  hp drat    wt  qsec vs gear carb transmission
## 1 21.0   6  160 110 3.90 2.620 16.46  0    4    4            M
## 2 21.0   6  160 110 3.90 2.875 17.02  0    4    4            M
## 3 22.8   4  108  93 3.85 2.320 18.61  1    4    1            M
## 4 21.4   6  258 110 3.08 3.215 19.44  1    3    1            A
## 5 18.7   8  360 175 3.15 3.440 17.02  0    3    2            A
## 6 18.1   6  225 105 2.76 3.460 20.22  1    3    1            A
```

From the boxplot we can see a significant difference between manual and automatic transmission, where we see a higher MPG for manual.

Let's see the Linear model for the outcome MPG with Transmission as the predictor

```
#Linear Model MPG ~ transmission
lmd1<-lm(mpg~transmission,mydata)
summary(lmd1)$coeff
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)   17.147368   1.124603 15.247492 1.133983e-15
## transmissionM  7.244939   1.764422  4.106127 2.850207e-04
```

```
summary(lmd1)$adj.r.squared
```

```
## [1] 0.3384589
```

We can see from the model that there's a significante increase for in MPG for the manual transmission. Since the value of the adjusted R for model 1 is of only 33,85%, we shoudl consider other variables that may produce better model

Model Analysis and Selection

Let's first check the correlation between all variables for the mtcars data (plot can be seen on Appendix - Figures)

From the correlation plot it seems that variable cyl, disp, hp and wt have the strongest correlation.

We'll create two new models to verify this.

1. Model with predictors transmission + cyl + disp + hp + wt

```
#Linear Model MPG ~ transmission + cyl + disp + hp + wt
lmd2<-lm(mpg~transmission + cyl + disp + hp + wt,mydata)
summary(lmd2)$coeff
```

```
##                  Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)   38.20279869 3.66909647 10.412045 9.084987e-11
## transmissionM  1.55649163 1.44053603  1.080495 2.898430e-01
## cyl           -1.10637984 0.67635506 -1.635797 1.139322e-01
## disp           0.01225708 0.01170645  1.047036 3.047194e-01
## hp            -0.02796002 0.01392172 -2.008374 5.509659e-02
## wt            -3.30262301 1.13364263 -2.913284 7.256888e-03
```

```
summary(lmd2)$adj.r.squared
```

```
## [1] 0.8272816
```

2. Model with all predictors

```
#Linear Model MPG ~ All variables
lmd3<-lm(mpg~.,mydata)
summary(lmd3)$coeff
```

```
##                  Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept)   12.30337416 18.71788443  0.6573058 0.51812440
## cyl           -0.11144048  1.04502336 -0.1066392 0.91608738
## disp           0.01333524  0.01785750  0.7467585 0.46348865
## hp            -0.02148212  0.02176858 -0.9868407 0.33495531
## drat           0.78711097  1.63537307  0.4813036 0.63527790
## wt            -3.71530393  1.89441430 -1.9611887 0.06325215
## qsec           0.82104075  0.73084480  1.1234133 0.27394127
## vs             0.31776281  2.10450861  0.1509915 0.88142347
## gear           0.65541302  1.49325996  0.4389142 0.66520643
## carb          -0.19941925  0.82875250 -0.2406258 0.81217871
## transmissionM  2.52022689  2.05665055  1.2254035 0.23398971
```

```
summary(lmd3)$adj.r.squared
```

```
## [1] 0.8066423
```

From the values for adjusted R in both new models it looks as the second model is the best, with 82% variance explained

Let's confirm with anova.

```
#lets now compare the 3 models
anova(lmd1,lmd2,lmd3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ transmission
## Model 2: mpg ~ transmission + cyl + disp + hp + wt
## Model 3: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + gear + carb +
##     transmission
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     26 163.12  4    557.78 19.8538 6.809e-07 ***
## 3     21 147.49  5     15.63  0.4449    0.8121
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion

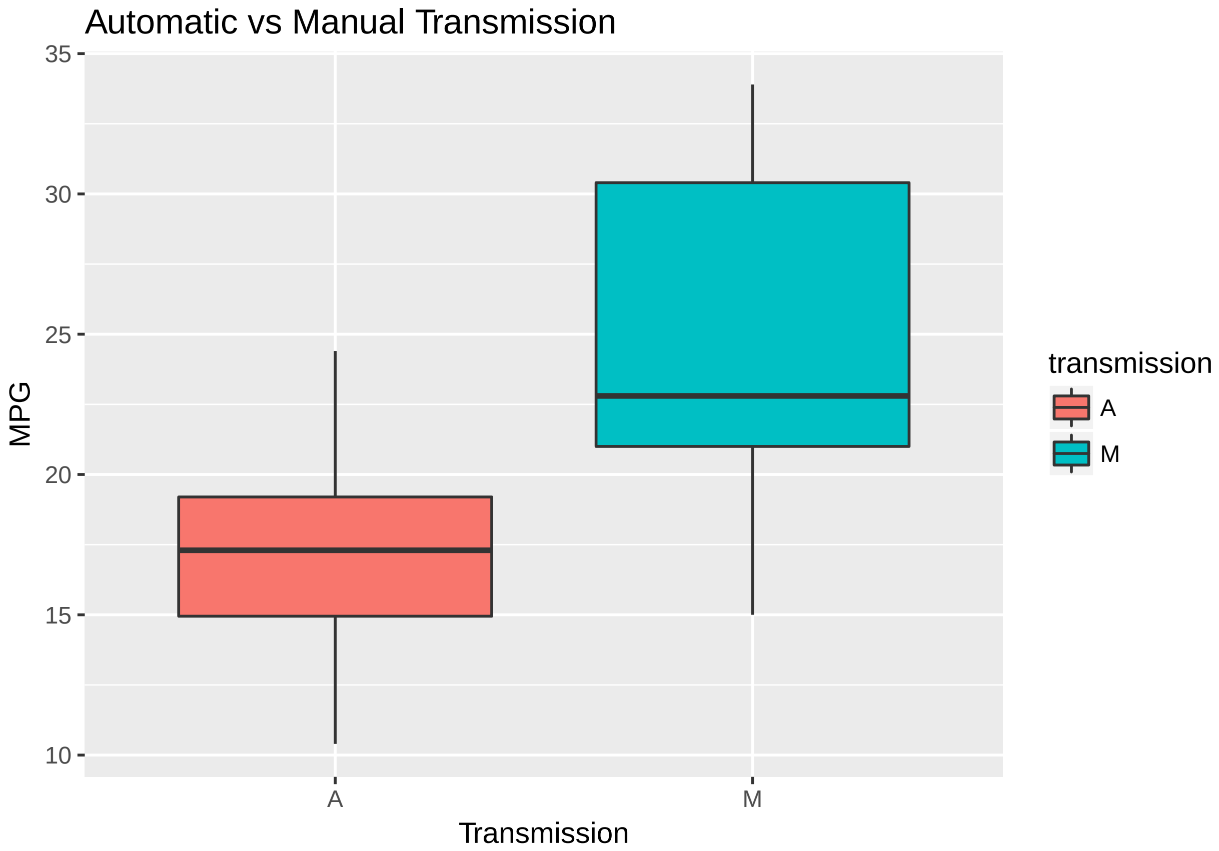
We clearly see that the second model is the best (MPG ~ Transmission + Cyl + Disp + HP + WT). Based on this model the difference between Auto and Manual Transmission is of 1.55 MPG (on Appendix - figures we can see the residuals for this model)

Appendix

Figures

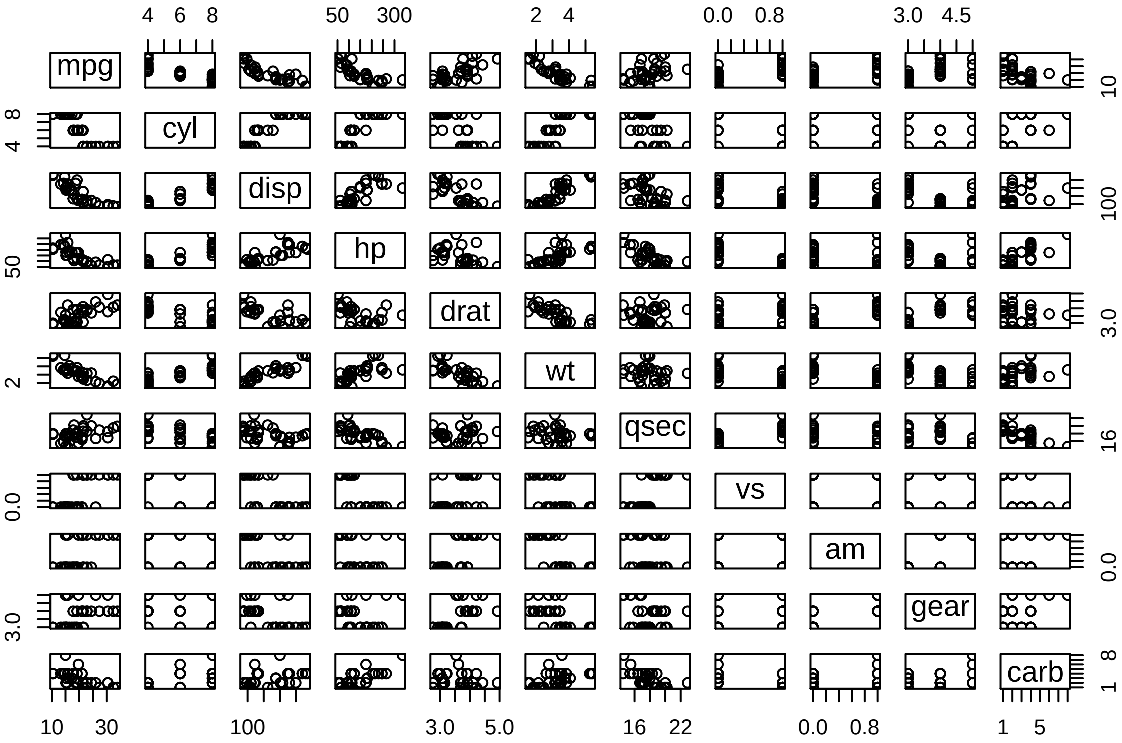
BoxPlot for data Mtcars

```
#plot data
plotdata
```



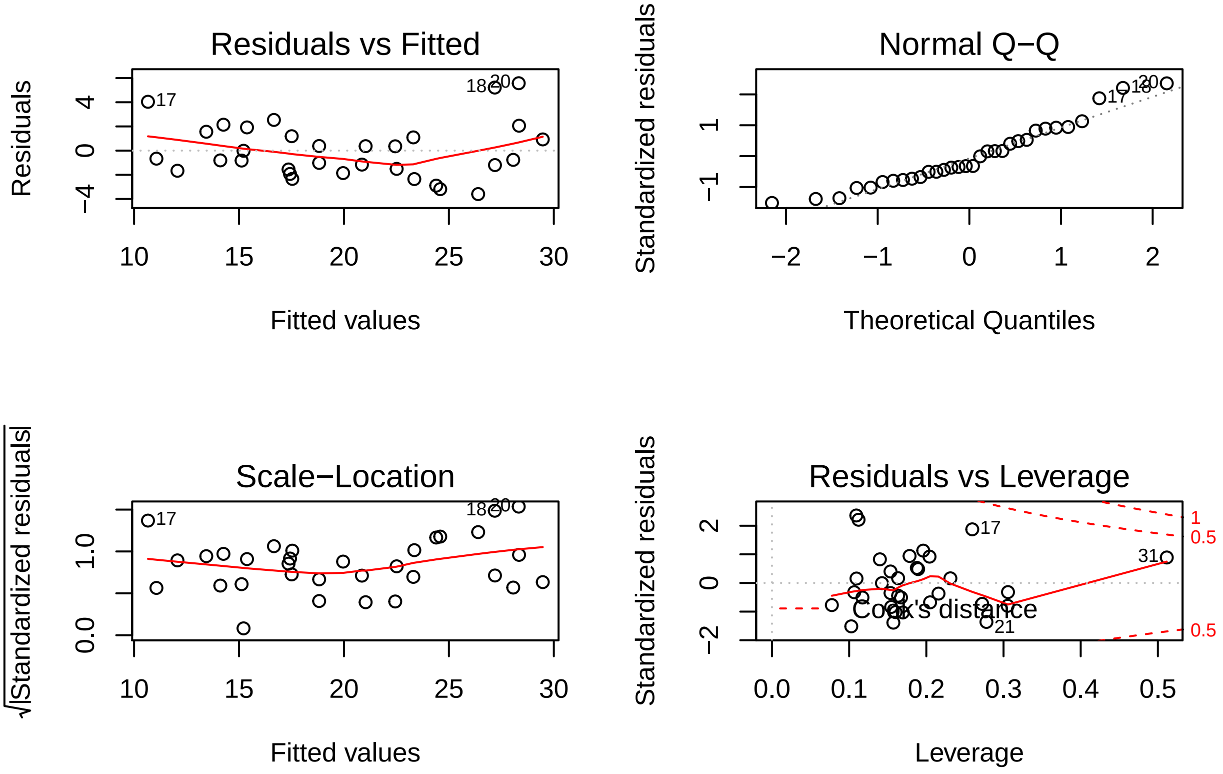
Data Correlation plot for data Mtcars

```r
#Display Data Correlation
pairs(mpg~ . ,data=mtcars)
```



Residuals for Model 2: MPG ~ Transmission + Cyl + Disp + HP + WT

```
#display residuals
par(mfrow=c(2,2))
plot(lmd2)
```


Residuals vs Fitted
Residuals
Fitted values
−4
0
4
10
15
20
25
30
17
18
20
Normal Q−Q
Standardized residuals
Theoretical Quantiles
−1
1
−2
−1
0
1
2
17
18
20
Scale−Location
√|Standardized residuals|
Fitted values
0.0
1.0
10
15
20
25
30
17
18
20
Residuals vs Leverage
Standardized residuals
Leverage
−2
0
2
0.0
0.1
0.2
0.3
0.4
0.5
17
31
21
Cook's distance
1
0.5
0.5