Regression Models Course Project

aps2201

August 31, 2017

Overview

In this project we want to see two things:

- Is an automatic or manual transmission better for MPG?
- The MPG difference between automatic and manual transmissions

First, we need to look at the `mtcars` dataset

```
summary(mtcars)
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
##        am              gear            carb      
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
##  Median :0.0000   Median :4.000   Median :2.000  
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812  
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000  
```

The vs and am are not supposed to be numeric, since they are actually codes for V/S and automatic/manual respectively.

Lets fix that

```
mtcars = mtcars %>%
  mutate(vs = ifelse(vs == "0","V","S"),am = ifelse(am == "0","automatic","manual")) %>%
  mutate(vs = factor(vs,levels=c("V","S")),am = factor(am,levels = c("automatic","manual")))
```

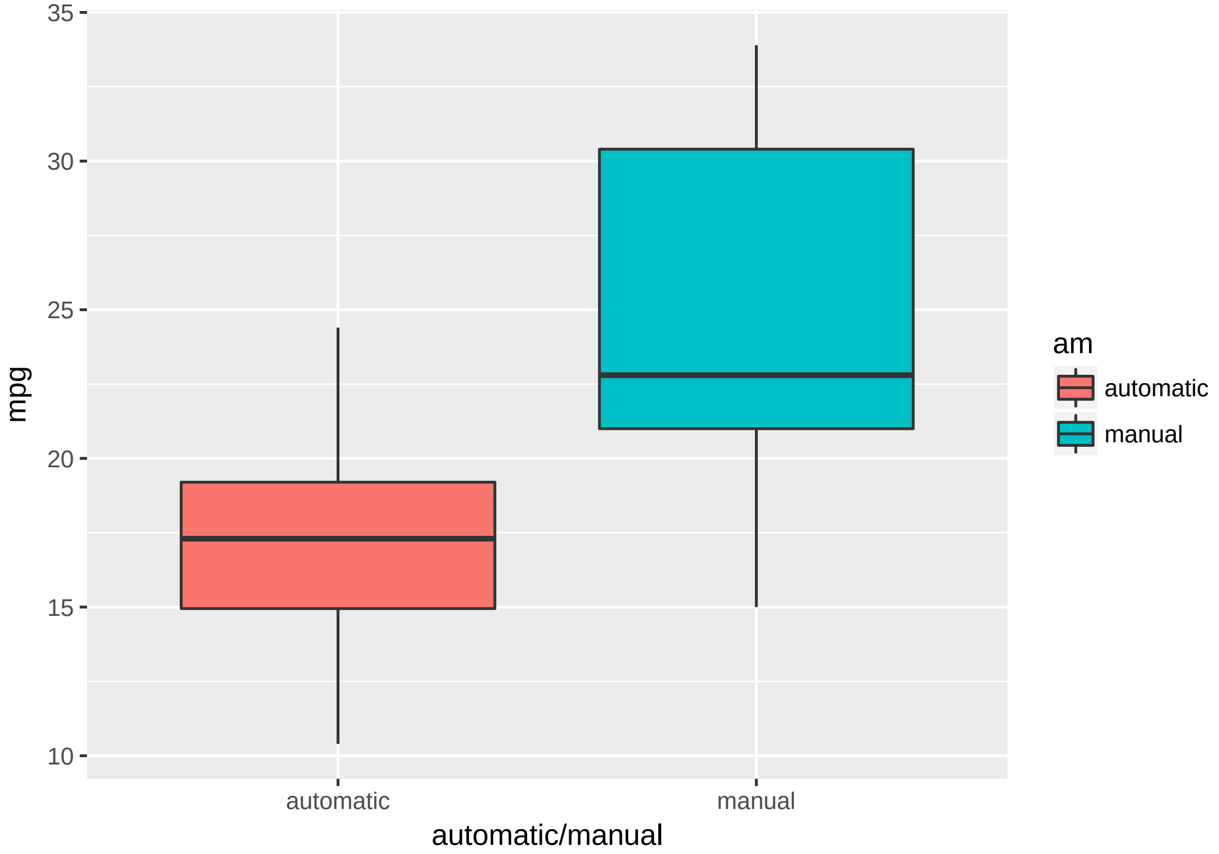
```
summary(mtcars)[,8:9]
```

```
##  vs             am    
##  V:18   automatic:19  
##  S:14   manual   :13  
##                       
##                       
##                       
```

```
##
```

ok, now we need to answer the question *Is an automatic or manual transmission better for MPG?*, to do this we can explore this by plotting the mpg with the transmission type.

```
amformpg
```



Now we can see that from miles per gallon (mpg) automatic seems to be the more gas guzzling compared to manual. Hence, manual is seemingly better compared to automatic regarding mpg.

Ok, this gives us a general idea, we need to look at the correlation table to see how it actually correlates, for this we need to convert the factors back to numeric for the `cor()` function to read.

```
##      [,1]  [,2]                 [,3]                 [,4]
## [1,] "mpg" "cyl"                "disp"               "hp"
## [2,] "1"   "-0.852161959426613" "-0.847551379262479" "-0.776168371826586"
##      [,5]                [,6]                 [,7]
## [1,] "drat"              "wt"                 "qsec"
## [2,] "0.681171907806749" "-0.867659376517228" "0.418684033921778"
##      [,8]                [,9]                [,10]
## [1,] "vs"                "am"                "gear"
## [2,] "0.664038919127593" "0.599832429454648" "0.480284757338842"
##      [,11]
## [1,] "carb"
## [2,] "-0.550925073902459"
```

By the looks of of the correlation table there are some outstanding numbers for the mpg correlations, we can figure that:

1. The lower the cylinder number the better its mpg.

2. The lower the displacement the better its mpg.
3. The lower the horse power the better its mpg.
4. The lower the weight the better its mpg.

automatic:

```
##       mpg             cyl             disp              hp
##  Min.   :10.40   Min.   :4.000   Min.   :120.1   Min.   : 62.0
##  1st Qu.:14.95   1st Qu.:6.000   1st Qu.:196.3   1st Qu.:116.5
##  Median :17.30   Median :8.000   Median :275.8   Median :175.0
##  Mean   :17.15   Mean   :6.947   Mean   :290.4   Mean   :160.3
##  3rd Qu.:19.20   3rd Qu.:8.000   3rd Qu.:360.0   3rd Qu.:192.5
##  Max.   :24.40   Max.   :8.000   Max.   :472.0   Max.   :245.0
##       drat              wt              qsec        vs              am
##  Min.   :2.760   Min.   :2.465   Min.   :15.41   V:12   automatic:19
##  1st Qu.:3.070   1st Qu.:3.438   1st Qu.:17.18   S: 7   manual   : 0
##  Median :3.150   Median :3.520   Median :17.82
##  Mean   :3.286   Mean   :3.769   Mean   :18.18
##  3rd Qu.:3.695   3rd Qu.:3.842   3rd Qu.:19.17
##  Max.   :3.920   Max.   :5.424   Max.   :22.90
##       gear            carb
##  Min.   :3.000   Min.   :1.000
##  1st Qu.:3.000   1st Qu.:2.000
##  Median :3.000   Median :3.000
##  Mean   :3.211   Mean   :2.737
##  3rd Qu.:3.000   3rd Qu.:4.000
##  Max.   :4.000   Max.   :4.000
```

manual:

```
##       mpg             cyl             disp              hp
##  Min.   :15.00   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
##  1st Qu.:21.00   1st Qu.:4.000   1st Qu.: 79.0   1st Qu.: 66.0
##  Median :22.80   Median :4.000   Median :120.3   Median :109.0
##  Mean   :24.39   Mean   :5.077   Mean   :143.5   Mean   :126.8
##  3rd Qu.:30.40   3rd Qu.:6.000   3rd Qu.:160.0   3rd Qu.:113.0
##  Max.   :33.90   Max.   :8.000   Max.   :351.0   Max.   :335.0
##       drat             wt              qsec        vs            am
##  Min.   :3.54   Min.   :1.513   Min.   :14.50   V:6   automatic: 0
##  1st Qu.:3.85   1st Qu.:1.935   1st Qu.:16.46   S:7   manual   :13
##  Median :4.08   Median :2.320   Median :17.02
##  Mean   :4.05   Mean   :2.411   Mean   :17.36
##  3rd Qu.:4.22   3rd Qu.:2.780   3rd Qu.:18.61
##  Max.   :4.93   Max.   :3.570   Max.   :19.90
##       gear            carb
##  Min.   :4.000   Min.   :1.000
##  1st Qu.:4.000   1st Qu.:1.000
##  Median :4.000   Median :2.000
##  Mean   :4.385   Mean   :2.923
##  3rd Qu.:5.000   3rd Qu.:4.000
##  Max.   :5.000   Max.   :8.000
```

So, now that we have proof on our assumption, we need to fit a reggression model to the correlation.

Pemember, we are just looking for mpg difference for automatic and manual (am), so we should build a basemodel that models the relation between those two variables. Here, we name them `basemodel` with `lm(mpg ~ am, data=mtcars)` as the model.

This is what it looks like:

```
basemodel
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Coefficients:
## (Intercept)           am
##      17.147        7.245
```

```
initialmodel
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Coefficients:
## (Intercept)          cyl         disp           hp         drat
##    12.30337     -0.11144      0.01334     -0.02148      0.78711
##          wt         qsec           vs           am         gear
##    -3.71530      0.82104      0.31776      2.52023      0.65541
##        carb
##    -0.19942
```

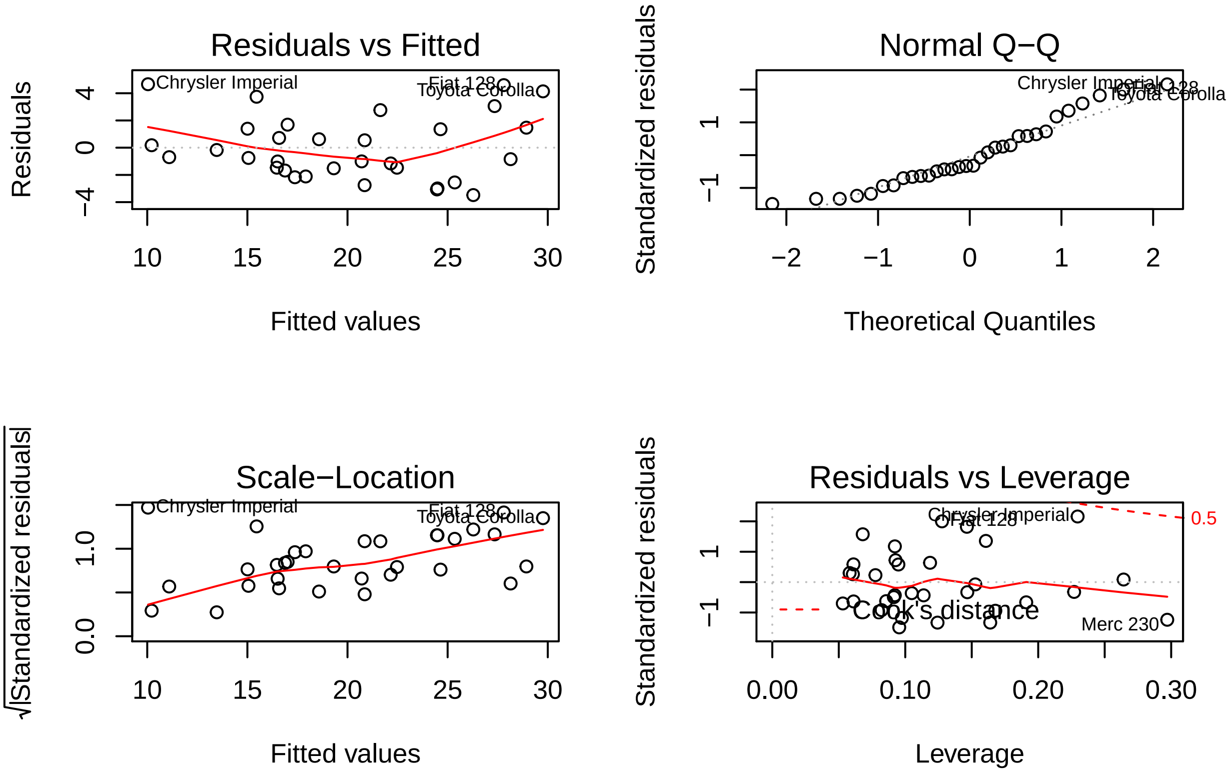
```
bestmodel
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Coefficients:
## (Intercept)           wt         qsec           am
##       9.618       -3.917        1.226        2.936
```

```
anova(basemodel, bestmodel)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + qsec + am
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     30 720.90
## 2     28 169.29  2    551.61 45.618 1.55e-09 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

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



```
t.test(mpg ~ am, data = mtcars)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## 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 in group automatic    mean in group manual
##                17.14737                24.39231
```