Regression Model Course Project

Himank Jain

July 12, 2019

##Executive Summary: This report is a course project within the Regression Models course on the Data Science Specialization by Johns Hopkins University on Coursera.

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?

```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
```

```r
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

##Exploratory Analysis: Let's look at the summary of mtcars mpg column:

```r
summary(mtcars$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   15.43   19.20   20.09   22.80   33.90
```

we see that the mean of mpg is 10.40.

Now let's separate automatic and manual transmission cars:

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

```
##   am      mpg
## 1  0 17.14737
## 2  1 24.39231
```

we note that mpg mean for automatic tranmission is 17.15 and for manual transmission it is 24.4 therefore, it seems at first that mean of mpg in manual trasmission cars is about 7.25 more than mean of mpg in automatic transmission.

##Linear Model: Let's fit a linear model with mpg as outcome and transmission mode as predictor:

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

```
##
## 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 ***
## am             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 that mean for automatic transmission is 17.15 and automatic transmission in 7.25 more. However the R squared value is 0.36 suggesting only 36% of model variance is explained by our model. Hence we need more predictors in it.

```r
fit2<-lm(data=mtcars,mpg~am+cyl)
fit3<-lm(data=mtcars,mpg~am+cyl+disp+hp)
fit4<-lm(data=mtcars,mpg~am+cyl+disp+hp+drat+wt)
fit5<-lm(data=mtcars,mpg~.)
anova(fit2,fit3,fit4,fit5)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am + cyl
## Model 2: mpg ~ am + cyl + disp + hp
## Model 3: mpg ~ am + cyl + disp + hp + drat + wt
## Model 4: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     29 271.36
```

```
## 2      27 216.37  2    54.995 3.9150 0.03589 *
## 3      25 162.43  2    53.934 3.8395 0.03792 *
## 4      21 147.49  4    14.939 0.5318 0.71378
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We note from p values in anova ,that till model fit4 the predictors are significant. Hence we use model fit4.

summary(fit4)

```
##
## Call:
## lm(formula = mpg ~ am + cyl + disp + hp + drat + wt, data = mtcars)
##
## Residuals:
##    Min     1Q Median    3Q    Max
## -3.437 -1.574 -0.688  1.310  5.551
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.04938    7.60553   4.740 7.31e-05 ***
## am           1.37506    1.56866   0.877  0.38906
## cyl         -1.03335    0.72405  -1.427  0.16590
## disp         0.01257    0.01195   1.052  0.30307
## hp          -0.02887    0.01444  -1.999  0.05658 .
## drat         0.48586    1.49495   0.325  0.74788
## wt          -3.27472    1.15685  -2.831  0.00903 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.549 on 25 degrees of freedom
## Multiple R-squared:  0.8557, Adjusted R-squared:  0.8211
## F-statistic: 24.72 on 6 and 25 DF,  p-value: 2.266e-09
```

from our model fit4 we note that the differnce in mean from automatic to manual transmission is 1.375. i.e much less than 7.25 we originally thought. This suggests that a lot of variance in mpg and transmission is explained by confounding variables like hp,wt,cyl etc.

##Conclusion:

####Is an automatic or manual transmission better for MPG? It appears that manual transmission cars are better for MPG compared to automatic cars. However when modeled with confounding variables like cyl,hp and wt, the difference is not as significant as it seemed to be in the beginning.

####Quantify the MPG difference between automatic and manual transmissions Initial Analysis shows that when only transmission was used in the model manual cars have an mpg increase of 7.245 than automatic. However, when confounding variables are included, the manual car advantage drops to 1.375.
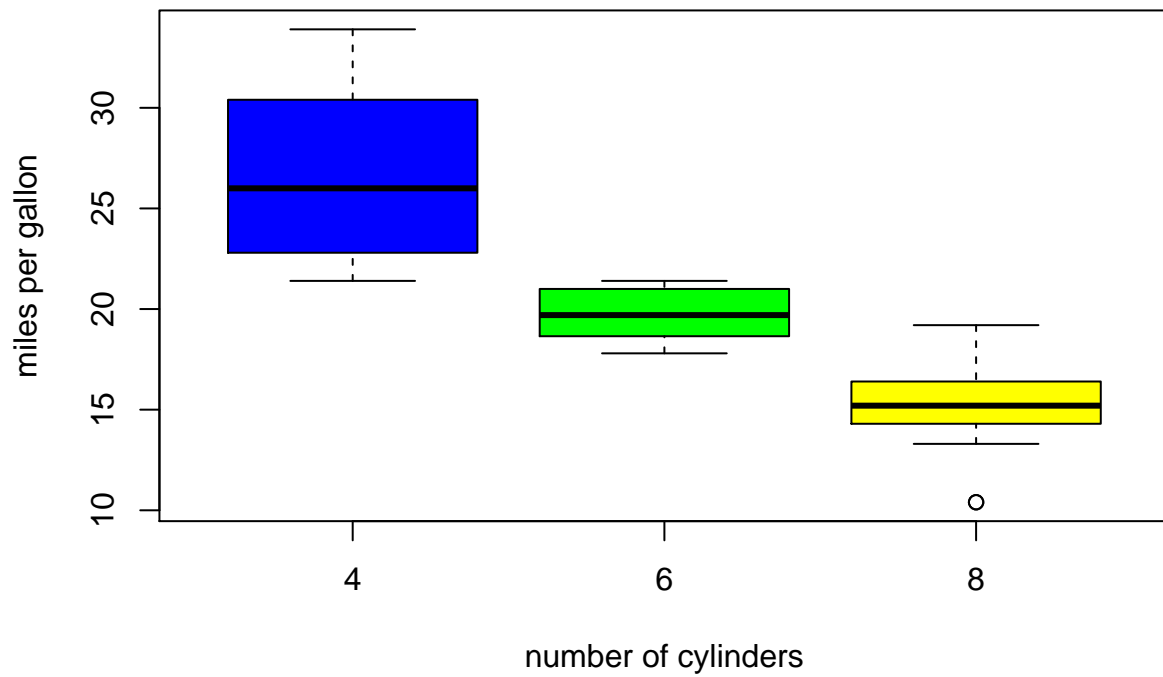
#Appendix:

boxplot(mpg ~ am,data=mtcars,col=c('purple','pink'),xlab="Transmission Automatic(0) vs Manual(0)",ylab=
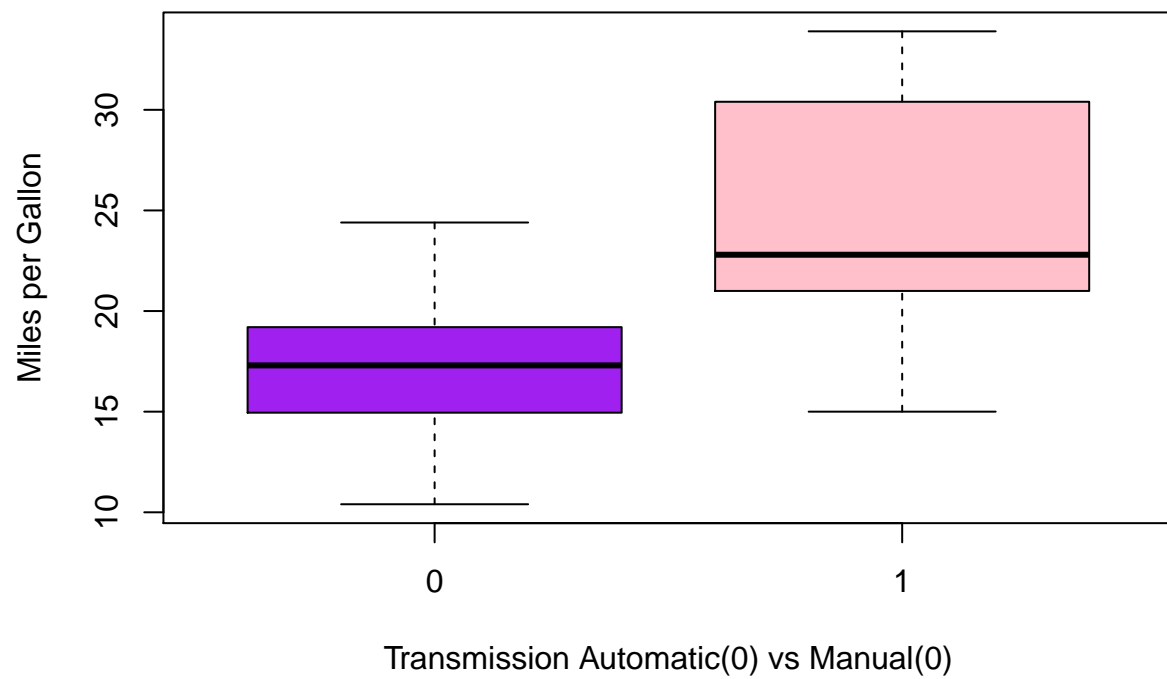
```
boxplot(mpg ~cyl, data=mtcars, col=(c("blue", "green", "yellow")), ylab="miles per gallon", xlab="number
```

Mileage by Cylinder

Scatter plot for all variables:

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