Regression Models Course Project

Author: **Fred Zhou**

In this document, we will try to answer the following questions:

- Q1: “Is an automatic or manual transmission better for MPG”
- Q2: “Quantify the MPG difference between automatic and manual transmissions”

By default, we assume that for the `mpg` , the lower the value the better.

(For `am` , `0` for automatic transmission, `1` for manual transmission.) ## summary of data

```
data("mtcars")
#Visulize the data first
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
```

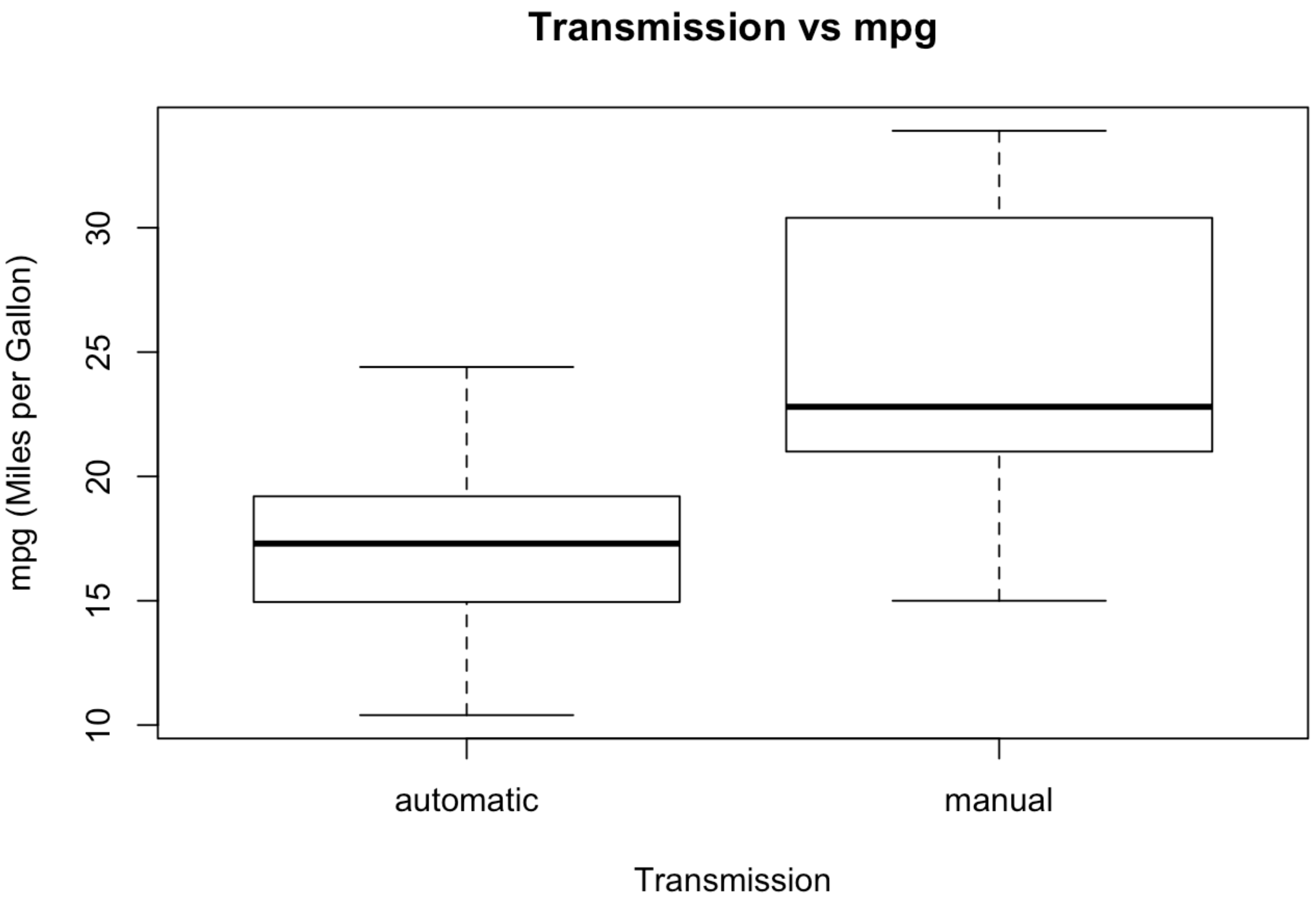
```
summary(mtcars)
```

```
##       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  
```

Q1. Is an automatic or manual transmission better for MPG

To answer this question, we assume that the all the variables in the population follow normal distribution. Thus we first use Student's T test to address wheter there's difference in these two groups

Visualize the data between AUTOMATIC and MANUAL



Student’s T-test between AUTOMATIC and MANUAL (alpha=0.05)

```
test_mpg=t.test(mtcars$mpg[mtcars$am==1],mtcars$mpg[mtcars$am==0])
print(test_mpg)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  mtcars$mpg[mtcars$am == 1] and mtcars$mpg[mtcars$am == 0]
## 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:
##   3.209684 11.280194
## sample estimates:
## mean of x mean of y 
##  24.39231  17.14737
```

```
print(paste('The P-value for the T-test between AUTOMATIC and MANUAL transmissions
for the mpg is ',round(test_mpg$p.value,digits = 4),sep=''))
```

```
## [1] "The P-value for the T-test between AUTOMATIC and MANUAL transmissions for
the mpg is 0.0014"
```

```
print(paste('Mean value for the mpg with AUTOMATIC transmissions:  ',round(test_mp
g$estimate[1],digits = 2),sep=''))
```

```
## [1] "Mean value for the mpg with AUTOMATIC transmissions:  24.39"
```

```
print(paste('Mean value for the mpg with MANUAL transmissions:  ',round(test_mpg$e
stimate[2],digits = 2),sep=''))
```

```
## [1] "Mean value for the mpg with MANUAL transmissions:  17.15"
```

Thus we could address that indeed the types of transmission will affect the `mpg` , and on average `AUTOMATIC` will bear a *higher consumption of fuel* against the `MANUAL` transmission, and the average difference is around *7.24* miles per Gallon used.

Q2. Quantify the MPG difference between automatic and manual transmissions

Correlation analysis winthin all variables against the mpg

```
sort(abs(cor(mtcars)[1,]))
```

```
##      qsec      gear      carb        am        vs      drat        hp
## 0.4186840 0.4802848 0.5509251 0.5998324 0.6640389 0.6811719 0.7761684
##      disp       cyl        wt       mpg
## 0.8475514 0.8521620 0.8676594 1.0000000
```

We already get the hint that the `AUTOMATIC/MANUAL` have impacts on the fuel consumption, thus from the correlation analsis we could guess that any variant with a higher correlation value against `AUTOMATIC/MANUAL` may contribute to the fuel consumption. including:

1. `vs` - V/S

2. `drat` - Rear axle ratio

3. `hp` - Gross horsepower

4. `disp` - Displacement (cu.in.)

5. `cyl` - Number of cylinders

6. `wt` - Weight (1000 lbs)

Thus, we could guess that it's reasonable to include any variable into the linear regressions. We could make a most general form of regression, then add in more variants to further optimize our model.

General model

We only take the `am` as variables to do the linear regression first:

```
fit_1 <- lm(mpg~am, data = mtcars)
summary(fit_1)
```

```
##
## 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
```

Based on the stat data we could address:

- On average, AUTOMATIC car have 17.15 MPG and MANUAL transmission cars have 7.25 MPG more
- The R^2 value is only 0.36, which means that our current model only explains 36% of the variance

lasso for the selection of variables

We try to include all the 7 possible variables (`am` , `vs` , `drat` , `hp` , `disp` , `cyl` , `wt`) meanwhile use Lasso to do the regression.

```
x<-model.matrix(mpg~am + vs + drat + hp + disp + cyl + wt,data=mtcars)
x=x[,-1]
glmnet1<-cv.glmnet(x=x,y=mtcars$mpg,type.measure='mse',nfolds=5,alpha=.5)
coef(glmnet1,s=9.8,exact=TRUE)
```

```
## 8 x 1 sparse Matrix of class "dgCMatrix"
##                       1
## (Intercept) 20.61147293
## am           .
## vs           .
## drat         .
## hp           .
## disp         .
## cyl         -0.01879497
## wt          -0.12574530
```

Based on the lasso results together with the correlation test, we could get the idea that the `wt cyl` affect most for the `mpg` .

Advanced model - linear regression using `wt cyl` and `am`

```
fit_2 <- lm(mpg ~ am + wt + cyl, data = mtcars)
summary(fit_2)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + cyl, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1735 -1.5340 -0.5386  1.5864  6.0812
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.4179     2.6415  14.923 7.42e-15 ***
## am            0.1765     1.3045   0.135  0.89334
## wt           -3.1251     0.9109  -3.431  0.00189 **
## cyl          -1.5102     0.4223  -3.576  0.00129 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.612 on 28 degrees of freedom
## Multiple R-squared:  0.8303, Adjusted R-squared:  0.8122
## F-statistic: 45.68 on 3 and 28 DF,  p-value: 6.51e-11
```

Based on the stat data we could address:

- `MANUAL` is slightly beneficial for the fuel saving, after model adjusting the value comes to be *0.1765* miles per gallon.
- `wt` and `cyl` affect huge against the `mpg` , which is appearant since more cylinders or more load will eventually consume more fuel.