Regression Course Project

Louis

10/17/2020

Coursera Regression Models Course Project

I load up the necessary libraries and the mtcars dataset

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)

data("mtcars")
```

I transform the data to make it easier to work with

```
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$gear <- as.factor(mtcars$gear)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- factor(mtcars$am, labels = c("Automatic", "Manual"))
```

I now explore the relationship between the miles per gallon(mpg) and the transmission type(am)

```
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
```

```
fit0 <- lm(mpg ~ am, mtcars)
summary(fit0)
```

```
##
## 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
```

From the looks of it, it seems like manual cars have on average 7.24494 more miles per gallon then automatic cars.

Also look at Figure 1

However it seems like I will need to make an multivariate model based on R squared being 36% meaning it is not entirely accurate.

I need to now do some data exploration in order to find which other variables are correlated with mpg.

```
vari_fit <- aov(mpg ~ . , data = mtcars)

summary(vari_fit)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## cyl          2  824.8   412.4  60.249 5.95e-09 ***
```

```
## disp         1   57.6    57.6   8.421  0.00914 **
## hp           1   18.5    18.5   2.703  0.11660
## drat         1   11.9    11.9   1.741  0.20273
## wt           1   55.8    55.8   8.150  0.01013 *
## qsec         1    1.5     1.5   0.223  0.64234
## vs           1    0.3     0.3   0.044  0.83584
## am           1   16.6    16.6   2.420  0.13627
## gear         2    5.0     2.5   0.367  0.69774
## carb         1    4.0     4.0   0.577  0.45677
## Residuals   19  130.1     6.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looks like cyl,disp,hp,drat,wt have high correlations with mpg.

Look at Figure 2.

We make new multivariate model now.

```
fit1 <- lm(mpg ~ am + cyl + wt + disp + hp + drat, data = mtcars)
```

We will now compare it with the original model(fit0) to determine if the additions to the model is necessary. This is done by using the anova funtion.

The null hypothesis is that we only need to use the original model(fit0) to see the effect of transmission type on miles per gallon.

```
anova(fit0,fit1)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl + wt + disp + hp + drat
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.9
## 2     24 150.1  6     570.8 15.211 3.944e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that the p-value is 3.944e-07, this means that we reject the null hypothesis, which means that the new model(fit1) is more correct in helping us to find the difference in miles per gallon for the different transmission types.

```
summary(fit1)
```

```
##
## Call:
```

```
## lm(formula = mpg ~ am + cyl + wt + disp + hp + drat, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8267 -1.4366 -0.4153  1.1649  5.0671
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 32.611986   6.274227   5.198 2.52e-05 ***
## amManual     1.681130   1.554386   1.082   0.2902
## cyl6        -3.026760   1.576680  -1.920   0.0669 .
## cyl8        -2.541967   3.059145  -0.831   0.4142
## wt          -2.726729   1.200207  -2.272   0.0323 *
## disp         0.004395   0.013090   0.336   0.7400
## hp          -0.033038   0.014476  -2.282   0.0316 *
## drat         0.326616   1.471086   0.222   0.8262
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.501 on 24 degrees of freedom
## Multiple R-squared:  0.8667, Adjusted R-squared:  0.8278
## F-statistic: 22.29 on 7 and 24 DF,  p-value: 4.768e-09
```

The Manual car now has only 1.68 miles per gallon more than the Automatic car.

This model has an 86.67% accuracy rate.

Appendix

Figure 1

```r
boxplot(mpg ~ am, data=mtcars, main = "Automatic vs Manual effect on Miles per Gallon", xlab = "Transmis
```

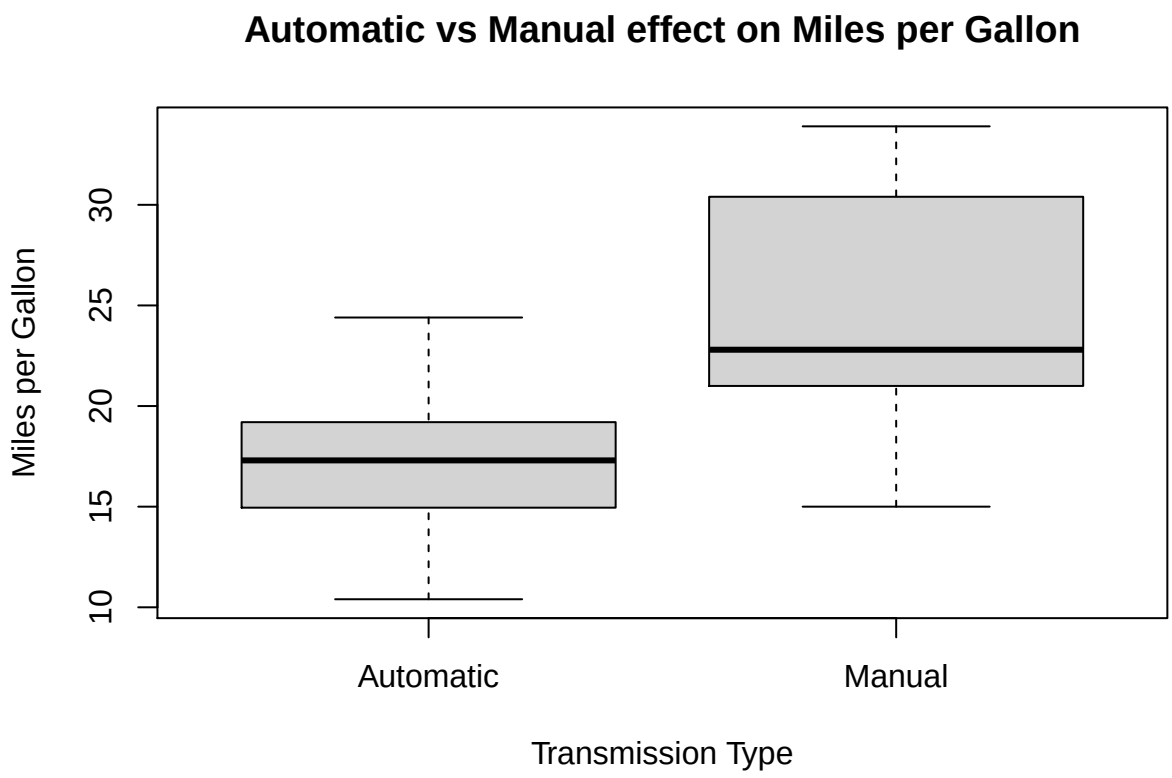
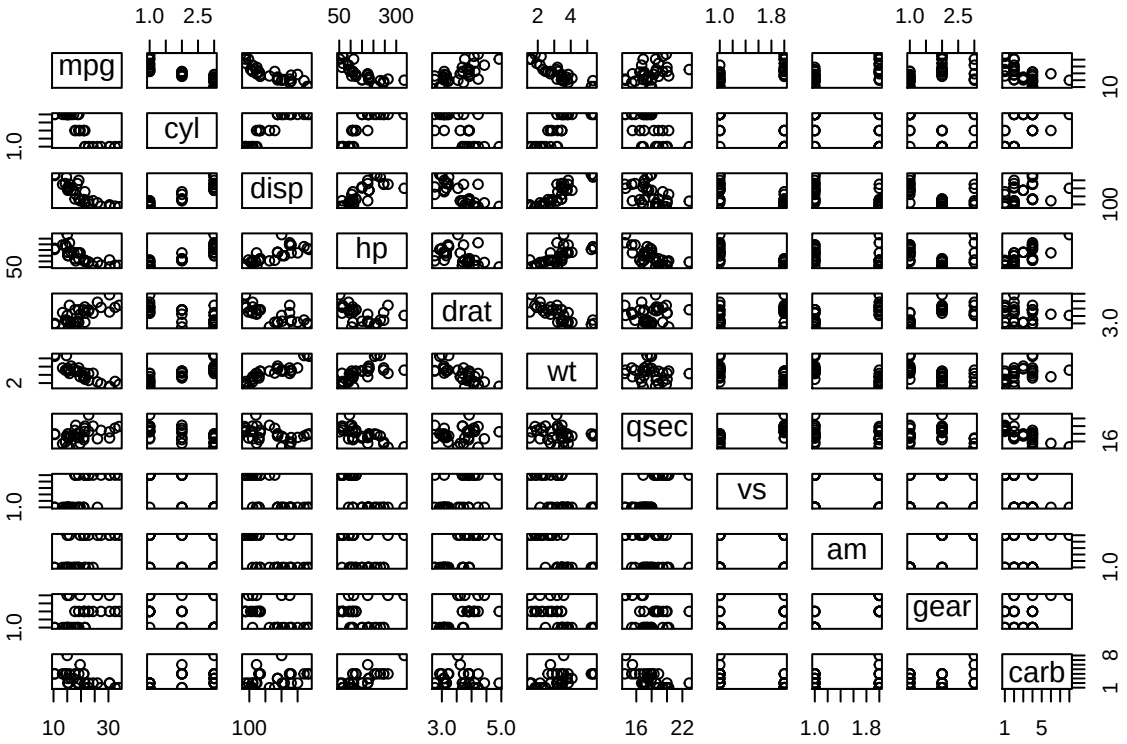


Figure 2

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


mpg
cyl
disp
hp
drat
wt
qsec
vs
am
gear
carb