WEEK 4 Project

Akash Gupta

29 May 2019

```
library(datasets)
data(mtcars)
```

Now , we will see the summary of our model

```
summary(lm(mpg~.-1,data=mtcars))$coef

##         Estimate Std. Error     t value   Pr(>|t|)
## cyl   0.35082641 0.76292423  0.45984438 0.65014009
## disp  0.01354278 0.01762273  0.76848373 0.45037109
## hp   -0.02054767 0.02143989 -0.95838513 0.34828334
## drat  1.24158213 1.46276742  0.84878985 0.40513967
## wt   -3.82613150 1.86238084 -2.05443023 0.05200271
## qsec  1.19139689 0.45942323  2.59324480 0.01659185
## vs    0.18972068 2.06824861  0.09173011 0.92774262
## am    2.83222230 1.97512820  1.43394353 0.16564985
## gear  1.05426253 1.34668717  0.78285629 0.44205756
## carb -0.26321386 0.81235653 -0.32401273 0.74898869
```

From summary we can see that Transmission has significant effcet on mpg

We also have to change the variable am to factor variable,since in the original dataset, this variable is numeric variable which is not what we want to analyze.

```
mtcars$am<-as.factor(mtcars$am)
```

mpg: Miles/(US) gallon by seeing the introduction of this dataset

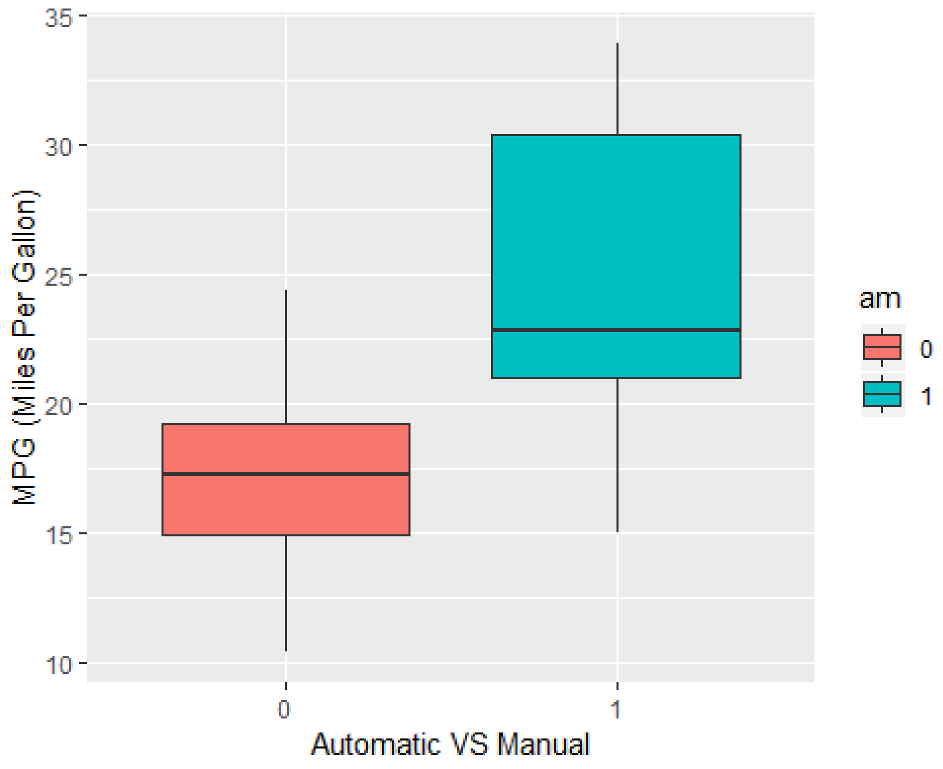
am : Is for distinguish the difference between automatic and manual transmissions

Answer to question 1:

```
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.5.2

g <- ggplot(data = mtcars, aes(y = mpg, x = am, fill = am))
g <- g + geom_boxplot()
g <- g + xlab("Automatic VS Manual") + ylab("MPG (Miles Per Gallon)")
g
```



As we can easily observe from our boxplot that Manual(1 in the boxplot) runs more efficient than Automatic(0 in the boxplot)

Now we can also show this effect from our simple linear model

```
fit1 <- lm(mpg ~ am - 1, data = mtcars)
summary(fit1)

##
## Call:
## lm(formula = mpg ~ am - 1, 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|)
## am0   17.147      1.125   15.25 1.13e-15 ***
## am1   24.392      1.360   17.94  < 2e-16 ***
## ---
## 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:  0.9487, Adjusted R-squared:  0.9452
## F-statistic: 277.2 on 2 and 30 DF,  p-value: < 2.2e-16
```

The summary shows that the variable am has a significant p-value, which is less than 5%, suggesting that the Null Hypothesis (there's no difference of the effect of Automatic and Manual on mpg) is rejected.

Answer to question 2

In this question we have to find the model which best explain the variable mpg

For we can use our ANOVA method by repeatedly adding different combination of regressor and find the bet of them ,for doing this we can use forward and backward stepwise slection

Forward stepwise selection starts with a intercept-only model, and then adds predictors to the model, one at the time, until all of the predictors are in the model. At each step the variable that gives the greatest additional improvement to the fit is added to the model. Backward method, on the other hand, begins with the full model, and then removes the least useful covariate, one at the time.

There are four models for.aic, for.bic, back.aic and back.bic to be considered, each of which is the best model with the largest R-squared is selected

```
fit.full <- lm(mpg ~ ., data = mtcars)
fit.am <- lm(mpg ~ am, data = mtcars)
for.aic <- step(lm(mpg ~ 1, data = mtcars), direction = "forward",
                scope = formula(fit.full), k = 2, trace = 0) # forward AIC
for.bic <- step(lm(mpg ~ 1, data=mtcars), direction = "forward",
                scope = formula(fit.full), k = log(32), trace = 0) # forward
BIC
back.aic <- step(fit.full, direction = "backward", k = 2, trace = 0) #
backward AIC
back.bic <- step(fit.full, direction = "backward", k = log(32), trace = 0) #
backward BIC
```

Now ,from R-squared value of all four model we can easily select which model is best

```
cbind(summary(for.aic)$r.squared,summary(for.bic)$r.squared,summary(back.aic)
$r.squared,summary(back.bic)$r.squared)

##          [,1]      [,2]      [,3]      [,4]
## [1,] 0.84315 0.8302274 0.8496636 0.8496636
```

The model back.bic is the best without any interaction of regresssor

Now we are going to check for interaction of regressors(predictors)

```
##summary(back.bic)

fit1<-lm(mpg~wt*qsec*am,data=mtcars)
fit2<-lm(mpg~wt+qsec*am,data=mtcars)
fit3<-lm(mpg~qsec+wt*am,data=mtcars)
```

```
fit4<-lm(mpg~am+wt*qsec,data=mtcars)

cbind((summary(fit1))$adj.r.squared,summary((fit2))$adj.r.squared,summary((fi
t3))$adj.r.squared,summary((fit4))$adj.r.squared)

##           [,1]      [,2]      [,3]      [,4]
## [1,] 0.8759496 0.8531624 0.8804219 0.8347545
```

Now from above analysis we can say the best model is fit3 i.e mpg = 9.723 + (1.017)qsec + (-2.937)wt + (14.079)am + (-4.141)wt*am

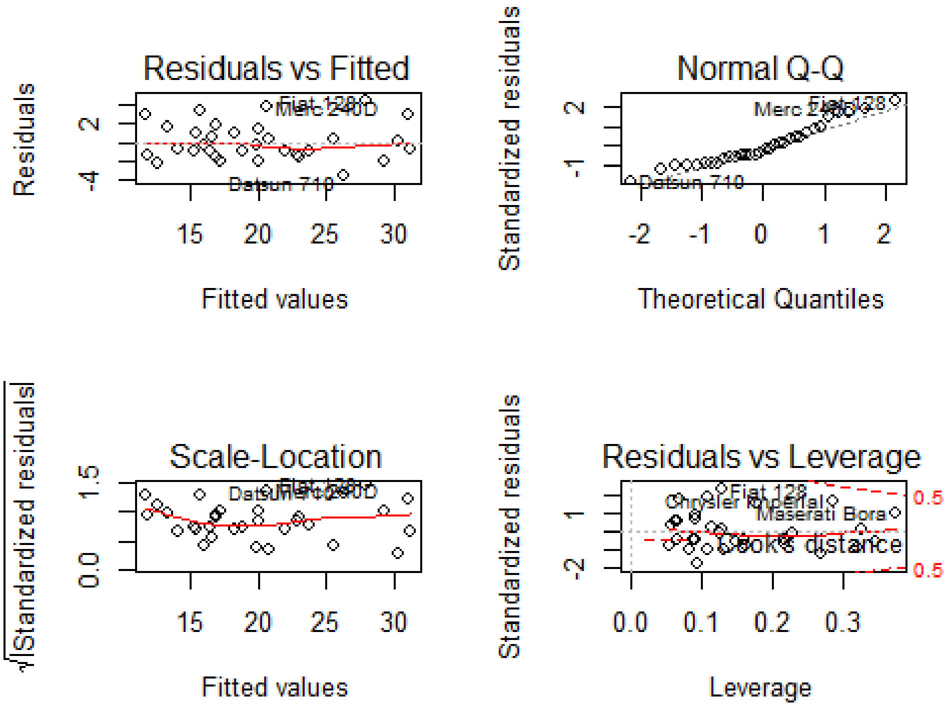
In term of uncertainty, the 95% confidence interval for the coefficients are shown below.

```
(confint(fit3))

##                  2.5 %    97.5 %
## (Intercept) -2.3807791 21.826884
## qsec         0.4998811  1.534066
## wt          -4.3031019 -1.569960
## am1          7.0308746 21.127981
## wt:am1      -6.5970316 -1.685721
```

Residual Diagnostics

```
par(mfrow=c(2,2))
plot(fit3)
```



Executive Summary

We first summarize the relationships between variables, and then fit a linear regression model that has the smallest BIC and the largest adjusted R2, followed by residual analysis and diagnostics. Wither with or wothout interaction, our model tells us a manual transmission is better for MPG, and no-pattern residual plots are indications for good model fitting.