S07_W4_Assignment - Building Regression Models with the mtcars Dataset

Instructions

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

1.Is an automatic or manual transmission better for MPG
2.Quantify the MPG difference between automatic and manual transmissions

Data

```
data("mtcars")
```

Analysis

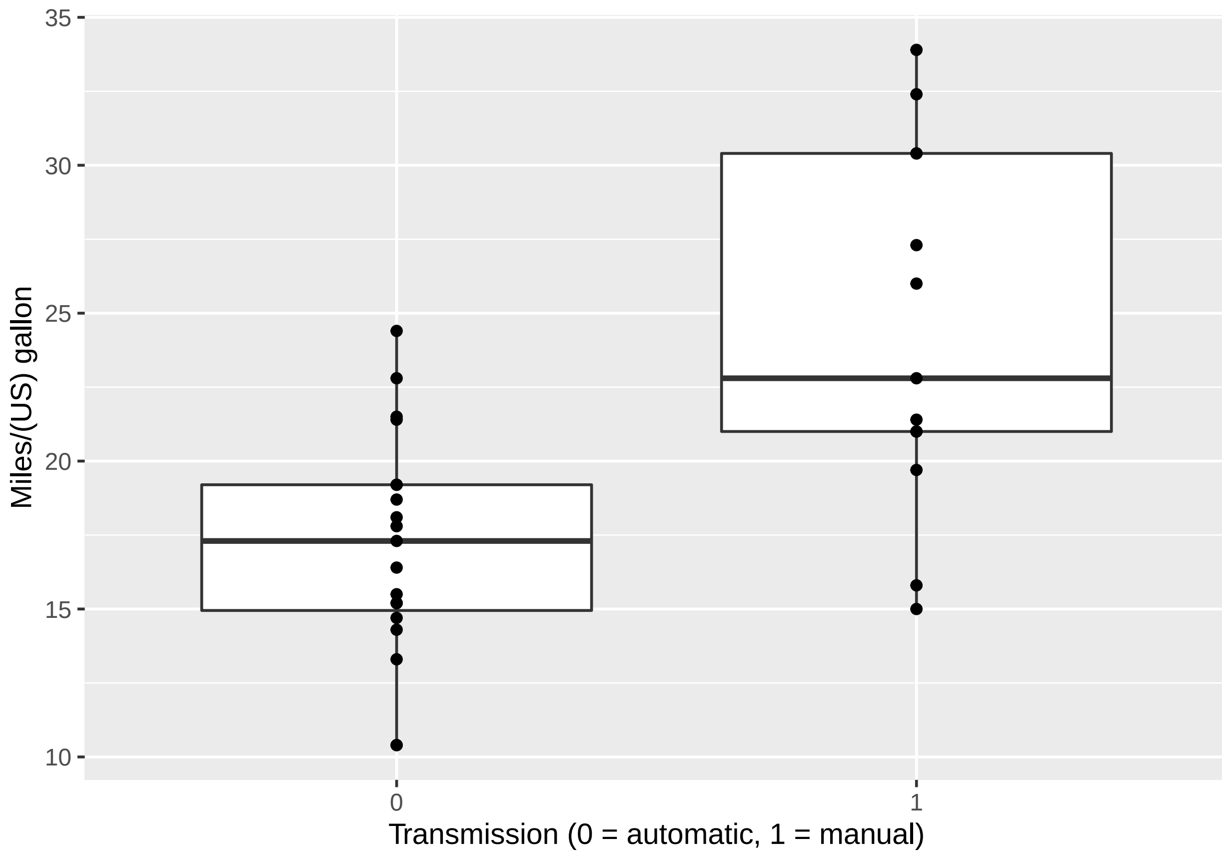
Environment Setup

```
library(ggplot2)
library(dbplyr)
library(tidyverse)
```

Exploratory Data Analyses

Since the main interest is to understand the relationship between transmission and MPG, isolate the two variables for exploratory data analysis

```
mpgAm <- mtcars %>% select(mpg, am) %>% mutate(am = as.factor(am))
ggplot(data = mpgAm, aes(x = am, y = mpg)) +
  geom_boxplot() + geom_point() +
  xlab('Transmission (0 = automatic, 1 = manual)') +
  ylab('Miles/(US) gallon')
```



From the box plot, it seems that manual transmission has higher mean of miles per gallon.

Fitting Models and Model Selection

First fit all variables to mpg and look at the diagnostics to decide which ones to remove (set type-I error at 5%)

```
raw <- mtcars %>% mutate(cyl = as.factor(cyl), vs = as.factor(vs), am = as.factor(am), gear = as.factor
fitAll <- lm(mpg ~ ., data = raw)
summary(fitAll)$coef[, 4]
```

```
## (Intercept)        cyl6        cyl8        disp          hp        drat
##  0.25252548  0.39746642  0.96317000  0.28267339  0.09393155  0.64073922
##          wt        qsec         vs1         am1       gear4       gear5
##  0.09461859  0.69966720  0.51150791  0.71131573  0.77332027  0.50889747
##       carb2       carb3       carb4       carb6       carb8
##  0.67865093  0.49546781  0.80956031  0.49381268  0.39948495
```

None of the coefficients has a p-value less than 5% in the full model, indicateing that variables should be selected - by slowly removing the most insignificant variables and refitting each time

```
which.max(summary(fitAll)$coef[, 4]) #the cyl variable (cyl8 is the least significant)
```

```
## cyl8
##    3
```

```r
fitRaw <- raw %>% select(-cyl); fitRm <- lm(mpg ~ ., data = fitRaw); summary(fitRm)$coef[, 4]
```

```
## (Intercept)        disp          hp        drat          wt        qsec 
##   0.4158127   0.2145504   0.1357694   0.2914041   0.1020825   0.5372086 
##         vs1         am1       gear4       gear5       carb2       carb3 
##   0.5622658   0.4964455   0.8004203   0.5903340   0.7423912   0.5839796 
##       carb4       carb6       carb8 
##   0.7337118   0.8632349   0.6856502
```

Again, there are no coefficients with a significant p-value after removing the cyl variable. The next most insignificant varible is removed and this process is repeated until all coefficients are significant

```r
#which.max(summary(fitRm)$coef[, 4]) #the carb variable
fitRaw <- fitRaw %>% select(-carb); fitRm <- lm(mpg ~ ., data = fitRaw)
#summary(fitRm)$coef[, 4]; which.max(summary(fitRm)$coef[, 4]) #the gear variable
fitRaw <- fitRaw %>% select(-gear); fitRm <- lm(mpg ~ ., data = fitRaw)
#summary(fitRm)$coef[, 4]; which.max(summary(fitRm)$coef[, 4]) #the vs variable
fitRaw <- fitRaw %>% select(-vs); fitRm <- lm(mpg ~ ., data = fitRaw)
#summary(fitRm)$coef[, 4]; which.max(summary(fitRm)$coef[, 4]) #the drat variable
fitRaw <- fitRaw %>% select(-drat); fitRm <- lm(mpg ~ ., data = fitRaw)
#summary(fitRm)$coef[, 4]; which.max(summary(fitRm)$coef[, 4]) #the disp variable
fitRaw <- fitRaw %>% select(-disp); fitRm <- lm(mpg ~ ., data = fitRaw)
#summary(fitRm)$coef[, 4]; which.max(summary(fitRm)$coef[, 4]) #the hp variable
fitRaw <- fitRaw %>% select(-hp); fitRm <- lm(mpg ~ ., data = fitRaw)
summary(fitRm)$coef[, 4]
```

```
##  (Intercept)           wt         qsec          am1 
## 1.779152e-01 6.952711e-06 2.161737e-04 4.671551e-02
```

Finally, after removing all the variables with insignificant p-values, three coefficients, wt (wieght of 1000 lbs), qseq (1/4 mile time), and am (Transmission), have p-values less than 0.05. The porperties of this model is further explored

```r
summary(fitRm)
```

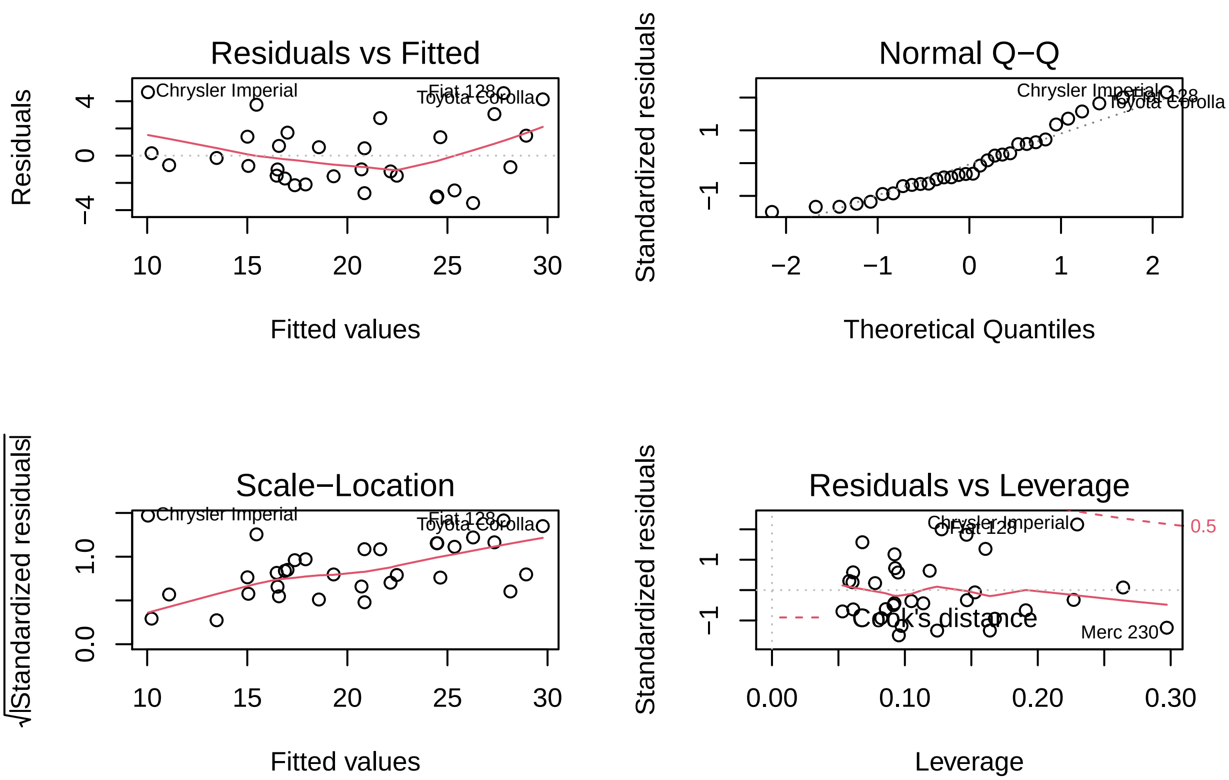
```
## 
## Call:
## lm(formula = mpg ~ ., data = fitRaw)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.4811 -1.5555 -0.7257  1.4110  4.6610 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   9.6178     6.9596   1.382 0.177915    
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am1           2.9358     1.4109   2.081 0.046716 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

This model indicates that given a fixed weight and 1/4 mile time, the mpg of an automatic is 9.6178 miles/gallon, but increases to 9.6178 + 2.9358 = 12.5536 miles/gallon for a manual. In addition, the adusted R-squared for this model is 0.8336, and the p-value for this model is 1.21e-11, indicating that we fail to reject the null hypothesis, and conclude that there is a significant relationship between the variables and mpg

Diagnostics

```r
par(mfrow = c(2, 2))
plot(fitRm)
```



The QQ plot shows a pretty good correlation of the standardized and theoretical residuals. There also doesn't seem to be any significant patterns in the other three plots, indicating a good fit of the selected model

Conclusions

Going back to the questions: understanding the relationship between transmission and mpg. From the model, we can conclude that when the weight and 1/4 mile time are the same for two cars, and one is an automatic and the other manual, the manual one will have an average of 2.9358 higher miles/gallon than the automatic car. Perhaps that's why a lot of race cars are manual?