Group_Assignment1

Tim Yang

2022-09-06

2.4 Exercises Problem 9.

Use R Markdown to generate a pdf file. Only one student from each group should submit the solution file.

This exercise involves the Auto data set studied in the R Videos. Make sure that the missing values have been removed from the data.

```
Auto = read.csv("/Users/hyeonwooyang/Desktop/Desktop/0_WUSTL/0_Business_Analytics/00_2022_Fall/4_DAT500S
Auto <- na.omit(Auto)
dim(Auto)
```

```
## [1] 392   9
```

(a) Which of the predictors are quantitative, and which are qualitative?

```
# View(Auto)
Auto$cylinders <- factor(Auto$cylinders)
Auto$year <- factor(Auto$year)
Auto$origin <- factor(Auto$origin)

summary(Auto)
```

```
##       mpg        cylinders  displacement     horsepower        weight    
##  Min.   : 9.00   3:  4     Min.   : 68.0   Min.   : 46.0   Min.   :1613  
##  1st Qu.:17.00   4:199     1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225  
##  Median :22.75   5:  3     Median :151.0   Median : 93.5   Median :2804  
##  Mean   :23.45   6: 83     Mean   :194.4   Mean   :104.5   Mean   :2978  
##  3rd Qu.:29.00   8:103     3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615  
##  Max.   :46.60             Max.   :455.0   Max.   :230.0   Max.   :5140  
##                                                                          
##   acceleration        year     origin      name          
##  Min.   : 8.00   73     : 40   1:245   Length:392        
##  1st Qu.:13.78   78     : 36   2: 68   Class :character  
##  Median :15.50   76     : 34   3: 79   Mode  :character  
##  Mean   :15.54   75     : 30                             
##  3rd Qu.:17.02   82     : 30                             
##  Max.   :24.80   70     : 29                             
##                  (Other):193                             
```

- **Quantitative predictors**: mpg, cylinders, displacement, horsepower, weight, acceleration
- **Qualitative predictors**: cylinders (factor), origin (factor), year (factor), name

(b) What is the range of each quantitative predictor? You can answer this using the range() function.

```
attach(Auto)
sapply(Auto[, -c(2, 7, 8, 9)], range)
```

```
##       mpg displacement horsepower weight acceleration
## [1,]  9.0           68         46   1613          8.0
## [2,] 46.6          455        230   5140         24.8
```

(c) What is the mean and standard deviation of each quantitative predictor?

```
sapply(Auto[, -c(2, 7, 8, 9)], mean)
```

```
##          mpg displacement   horsepower       weight acceleration 
##     23.44592    194.41199    104.46939   2977.58418     15.54133
```

```
sapply(Auto[, -c(2, 7, 8, 9)], sd)
```

```
##          mpg displacement   horsepower       weight acceleration 
##     7.805007   104.644004    38.491160   849.402560     2.758864
```

(d) Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

```
Auto_subset <- Auto[-c(10:85), ]
dim(Auto_subset)
```

```
## [1] 316   9
```

```
detach(Auto)
attach(Auto_subset)
sapply(Auto_subset[, -c(2, 7, 8, 9)], range)
```

```
##       mpg displacement horsepower weight acceleration
## [1,] 11.0           68         46   1649          8.5
## [2,] 46.6          455        230   4997         24.8
```

```
sapply(Auto_subset[, -c(2, 7, 8, 9)], mean)
```

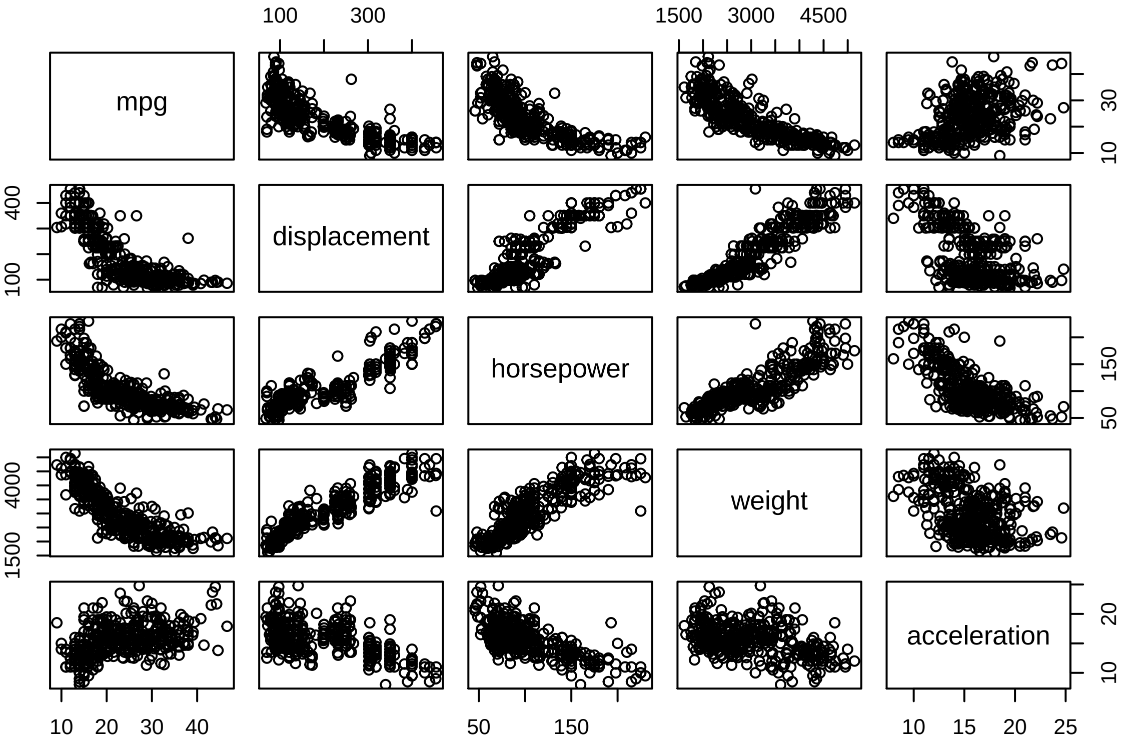
```
##          mpg displacement   horsepower       weight acceleration 
##     24.40443    187.24051    100.72152   2935.97152     15.72690
```

```
sapply(Auto_subset[, -c(2, 7, 8, 9)], sd)
```

```
##          mpg displacement   horsepower       weight acceleration 
##     7.867283    99.678367    35.708853   811.300208     2.693721
```

(e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

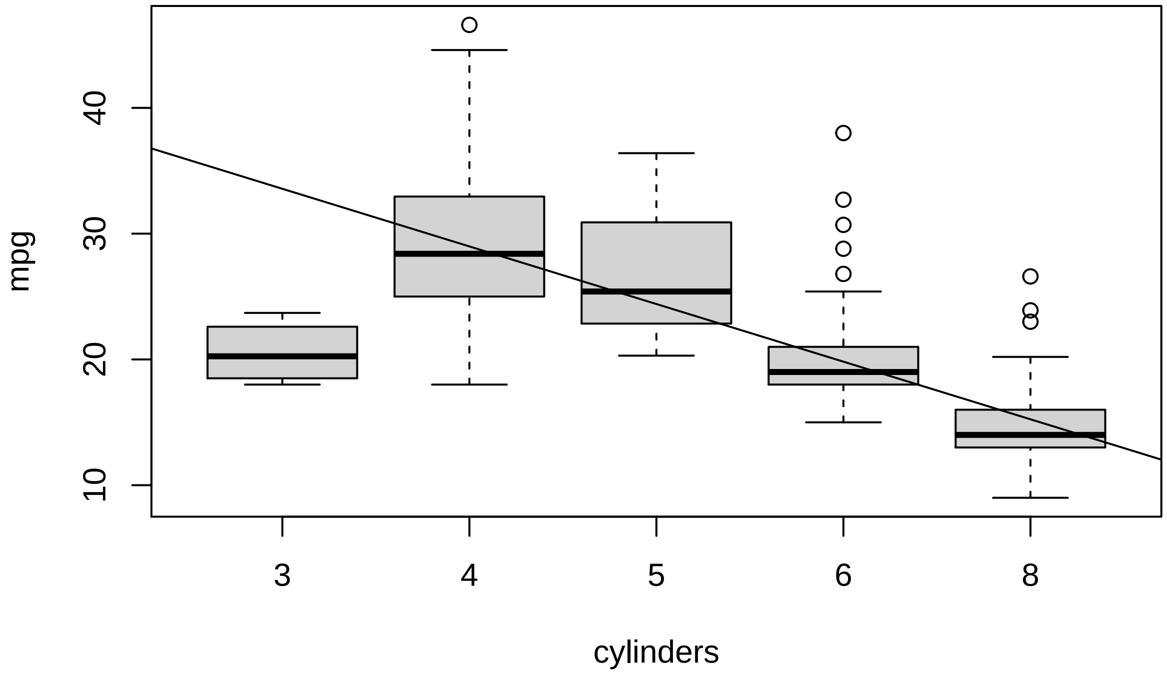
```
detach(Auto_subset)
attach(Auto)
pairs(~ mpg + displacement + horsepower + weight + acceleration, data = Auto)
```



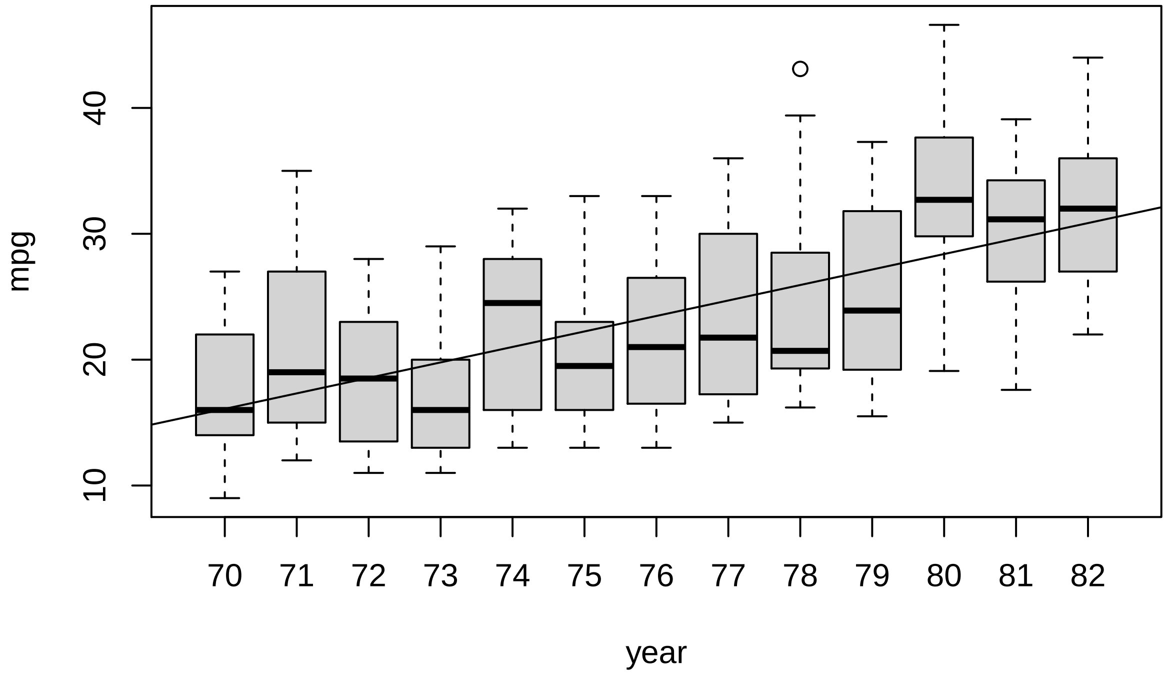
(f) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.

- displacement, horsepower, and weight variables are negatively correlated with the mpg variable, as shown in the plot above
- cylinder is negatively correlated with the mpg, while year and origin are positively correlated with the mpg, as shown in the graphs below

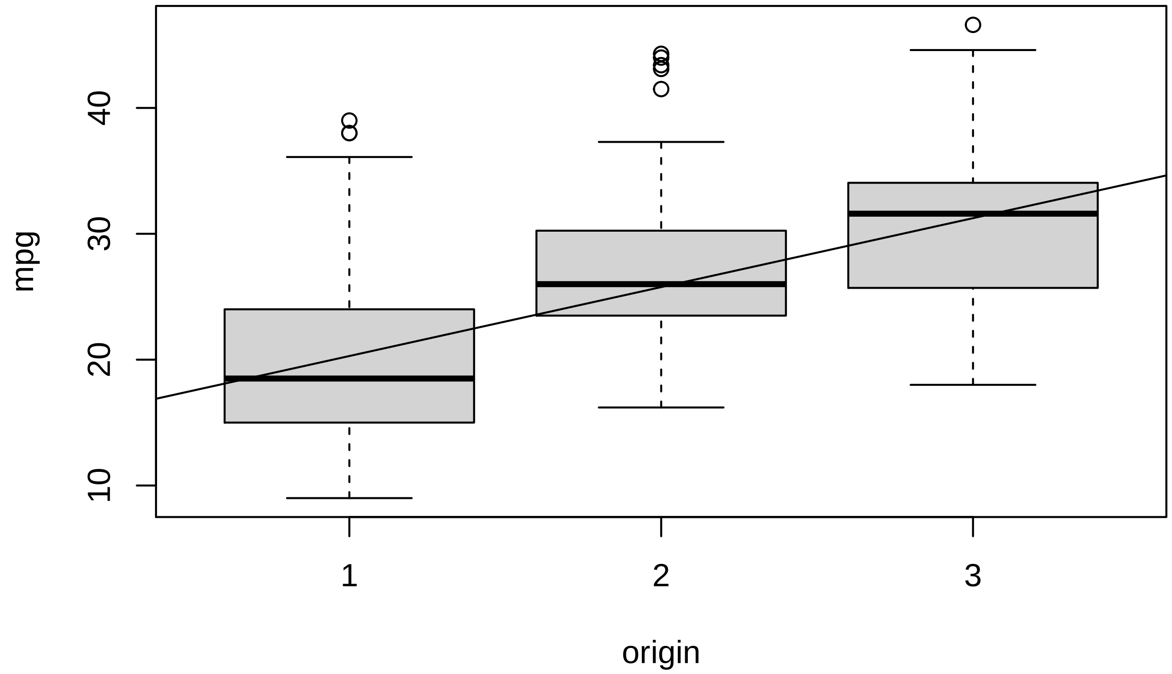
```
boxplot(mpg ~ cylinders)
regline <- lm(mpg ~ as.numeric(cylinders), data = Auto)
abline(regline)
```



```
boxplot(mpg ~ year)
regline <- lm(mpg ~ as.numeric(year), data = Auto)
abline(regline)
```



```
boxplot(mpg ~ origin)
regline <- lm(mpg ~ as.numeric(origin), data = Auto)
abline(regline)
```



Reference:

https://stat.ethz.ch/pipermail/r-help/2011-April/273755.html