Module 1: Introduction to Machine Learning

Seongryung Kim

2024-11-12

#2.4 Exercises Problem 9 ##This exercise involves the Auto data set studied in the R Videos. Make sure that the missing values have been removed from the data.

```
#Load necessary libraries
library(ISLR)
library(MASS)

#Load the Auto dataset
#Remove missing values first
Auto=read.csv("Auto.csv",header=T, na.strings = "?")
Auto =na.omit(Auto)
```

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

Answer - Quantitive: mpg, displacement, horsepower, weight, acceleration, year - Qualitative: cylinders, origin, name

```
sapply(Auto, class)
```

```
##          mpg    cylinders displacement   horsepower       weight accelerat
ion
##    "numeric"    "integer"    "numeric"    "integer"    "integer"    "numer
ic"
##         year       origin         name
##    "integer"    "integer"  "character"
```

```
head(Auto)
```

```
##   mpg cylinders displacement horsepower weight acceleration year origin
## 1  18         8          307        130   3504         12.0   70      1
## 2  15         8          350        165   3693         11.5   70      1
## 3  18         8          318        150   3436         11.0   70      1
## 4  16         8          304        150   3433         12.0   70      1
## 5  17         8          302        140   3449         10.5   70      1
## 6  15         8          429        198   4341         10.0   70      1
##                       name
## 1 chevrolet chevelle malibu
## 2         buick skylark 320
## 3        plymouth satellite
## 4             amc rebel sst
## 5               ford torino
## 6          ford galaxie 500
```

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

```
# Find the range of each quantitative predictor
ranges <- sapply(Auto[, c("mpg", "cylinders", "displacement", "horsepower", "
weight", "acceleration", "year")], range)
ranges
```

```
##       mpg cylinders displacement horsepower weight acceleration year
## [1,]  9.0         3           68         46   1613          8.0   70
## [2,] 46.6         8          455        230   5140         24.8   82
```

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

```
# Calculate mean and standard deviation
means <- sapply(Auto[, c("mpg", "cylinders", "displacement", "horsepower", "w
eight", "acceleration", "year")], mean)
sds <- sapply(Auto[, c("mpg", "cylinders", "displacement", "horsepower", "wei
ght", "acceleration", "year")], sd)

# Combine results into a table
summary_stats <- data.frame(Mean = means, SD = sds)
summary_stats
```

```
##                     Mean         SD
## mpg            23.445918   7.805007
## cylinders       5.471939   1.705783
## displacement  194.411990 104.644004
## horsepower    104.469388  38.491160
## weight       2977.584184 849.402560
## acceleration   15.541327   2.758864
## year           75.979592   3.683737
```

(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?

```
# Remove the 10th through 85th observations
Auto_subset <- Auto[-(10:85), ]

# Compute the range, mean, and standard deviation for the subset data
subset_ranges <- sapply(Auto_subset[, c("mpg", "cylinders", "displacement", "
horsepower", "weight", "acceleration", "year")], range)
subset_means <- sapply(Auto_subset[, c("mpg", "cylinders", "displacement", "h
orsepower", "weight", "acceleration", "year")], mean)
subset_sds <- sapply(Auto_subset[, c("mpg", "cylinders", "displacement", "hor
sepower", "weight", "acceleration", "year")], sd)
```

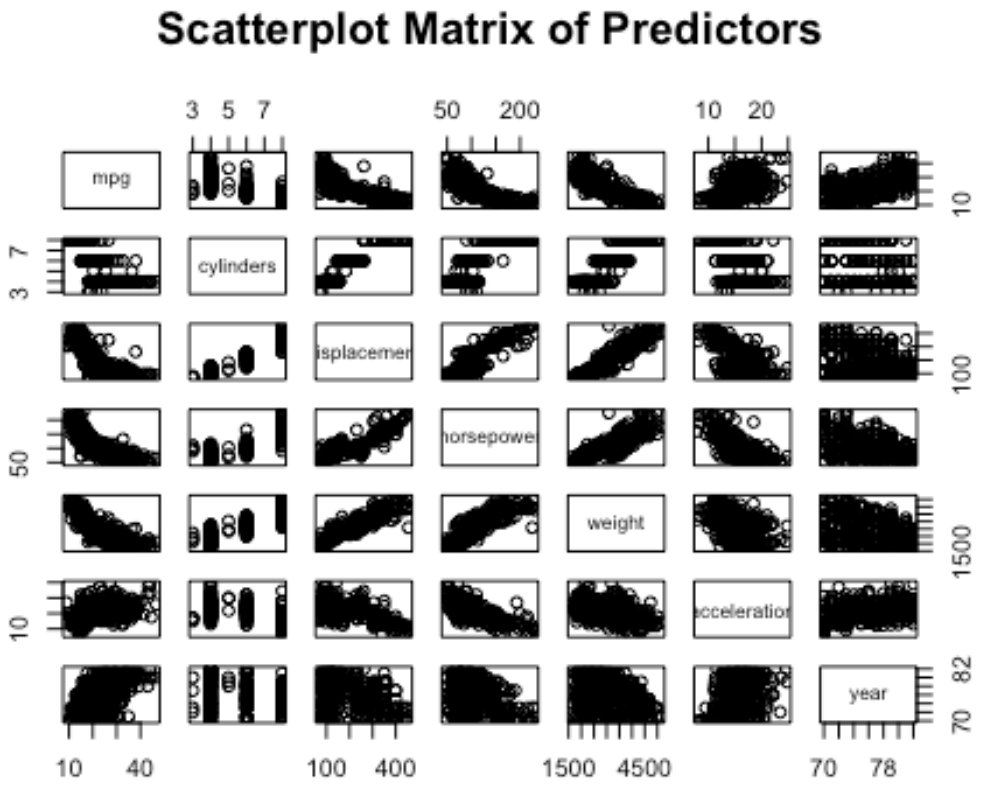
```r
# Combine results into a table
subset_stats <- data.frame(Range = subset_ranges[1, ], Mean = subset_means, S
D = subset_sds)
subset_stats
```

```
##                Range        Mean          SD
## mpg             11.0   24.404430   7.867283
## cylinders        3.0    5.373418   1.654179
## displacement    68.0  187.240506  99.678367
## horsepower      46.0  100.721519  35.708853
## weight        1649.0 2935.971519 811.300208
## acceleration     8.5   15.726899   2.693721
## year            70.0   77.145570   3.106217
```

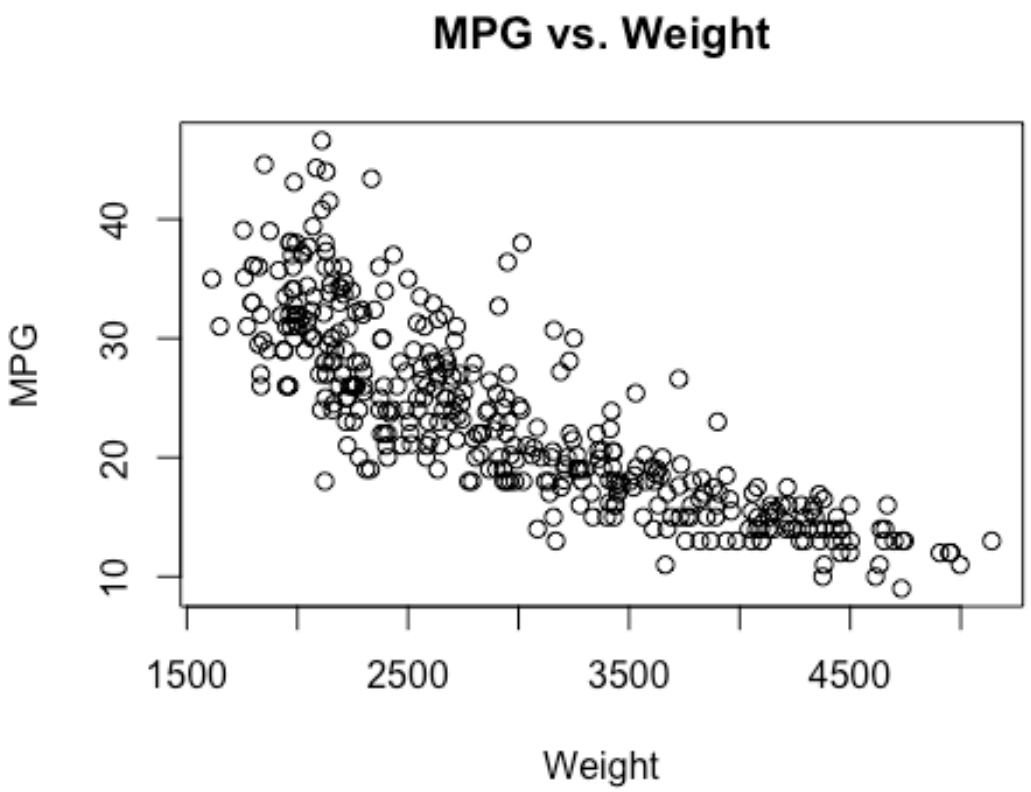
(e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

Answer - mpg has a negative correlation with weight and displacement, suggesting that cars with more weight and displacement tend to have lower fuel efficiency.

```r
# Scatterplot matrix for a quick overview of relationships
pairs(Auto[, c("mpg", "cylinders", "displacement", "horsepower", "weight", "a
cceleration", "year")], main = "Scatterplot Matrix of Predictors")
```



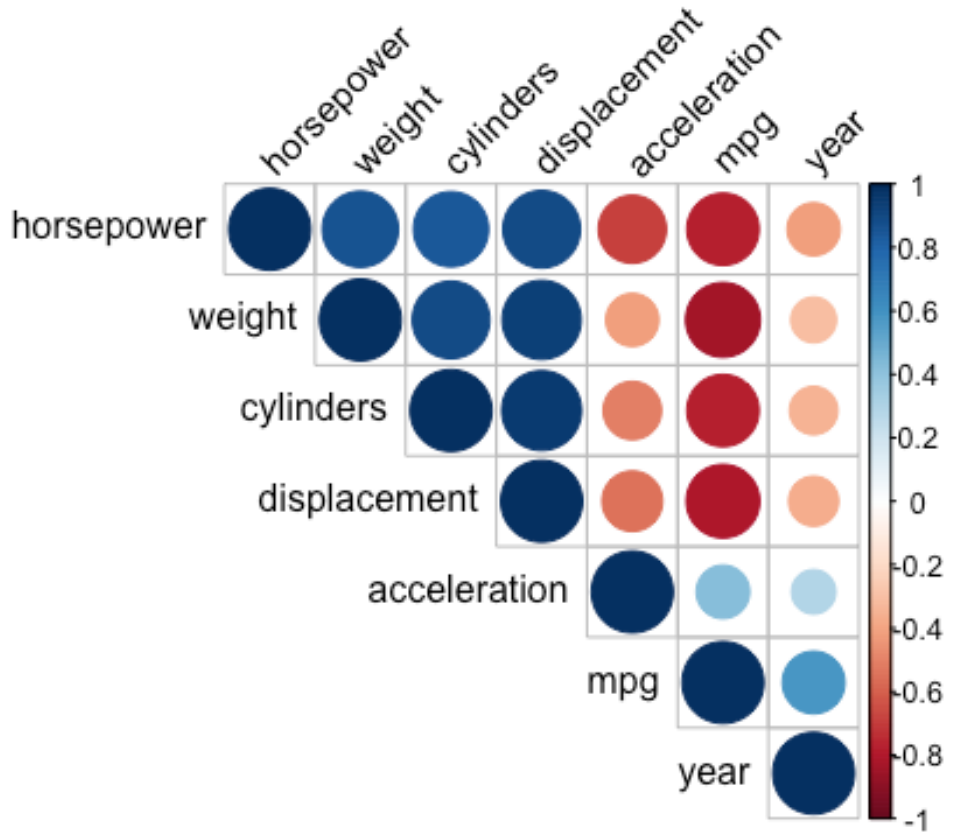

```r
# Individual scatterplot for mpg vs. weight
plot(Auto$weight, Auto$mpg,
     xlab = "Weight", ylab = "MPG",
     main = "MPG vs. Weight")
```




```r
# Correlation plot
library(corrplot)

## corrplot 0.95 loaded

cor_matrix <- cor(Auto[, c("mpg", "cylinders", "displacement", "horsepower",
"weight", "acceleration", "year")])
corrplot(cor_matrix, method = "circle", type = "upper", order = "hclust", tl.
col = "black", tl.srt = 45)
```




```
round(cor(Auto[,-9]),digits=3)
```

```
##                 mpg cylinders displacement horsepower weight acceleration
## mpg           1.000    -0.778       -0.805     -0.778 -0.832        0.423
## cylinders    -0.778     1.000        0.951      0.843  0.898       -0.505
## displacement -0.805     0.951        1.000      0.897  0.933       -0.544
## horsepower   -0.778     0.843        0.897      1.000  0.865       -0.689
## weight       -0.832     0.898        0.933      0.865  1.000       -0.417
## acceleration  0.423    -0.505       -0.544     -0.689 -0.417        1.000
## year          0.581    -0.346       -0.370     -0.416 -0.309        0.290
## origin        0.565    -0.569       -0.615     -0.455 -0.585        0.213
##                year origin
## mpg           0.581  0.565
## cylinders    -0.346 -0.569
## displacement -0.370 -0.615
## horsepower   -0.416 -0.455
## weight       -0.309 -0.585
## acceleration  0.290  0.213
## year          1.000  0.182
## origin        0.182  1.000
```

(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.

Answer - Weight: There is a negative relationship with mpg, indicating that heavier cars tend to have lower fuel efficiency. - Horsepower: This also shows some negative correlation with mpg, implying that cars with more horsepower tend to consume more fuel. - Displacement: Like weight and horsepower, this is negatively correlated with mpg.