MATH 189 HW4

Zijian Su
Zelong Zhou
Xiangyi Lin

Last Updated: February 10, 2023

Concrete contributions

All problems were done by Zijian Su, Zelong Zhou, Xiangyi Lin. All contributing equally to this assignment. Everyone put in enough effort.

Overview

In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the Auto data set. This data can be found in the ISLR package.

Packages

```
#install.packages("rmarkdown")
#install.packages('ISLR')
#install.packages('tools')
#install.packages('dplyr')
library("ISLR")
```

```
## Warning: package 'ISLR' was built under R version 4.1.3
```

```
data("Auto")
#tinytex::install_tinytex()
```

Question 1

Create a binary variable, mgp01, that contains 1 if mpg contains a value above its median, and a 0 if mpg contains a value below its median. You can compute the median using the median() function. Note you may find it helpful to use the data.frame() function to create a single data set containing both mpg01 and the other Auto variables.

Answer:

```
median_mpg <- median(Auto$mpg) #get the median of mpg
mpg01 <- c()               #if < median , put 0, else put 1
for (i in Auto$mpg){
  if (i < median_mpg){
    mpg01 <- append(mpg01,0)
  }
  else{
    mpg01 <- append(mpg01,1)
  }
}
Auto$mpg01 <- mpg01       # add a column in to dataset

#print out some sample data
cat("median mpg:", median_mpg,"\nindex  mpg, mpg01\n")
```

```
## median mpg: 22.75
## index  mpg, mpg01
```

```
for (i in 10 :15){
  cat(paste("",i,"   ",Auto$mpg[i],"  ", Auto$mpg01[i]),"\n")
}
```

```
##  10     15    0
##  11     15    0
##  12     14    0
##  13     15    0
##  14     14    0
##  15     24    1
```

mgp01 has been added as a new column of data to the Auto data.

Question 2

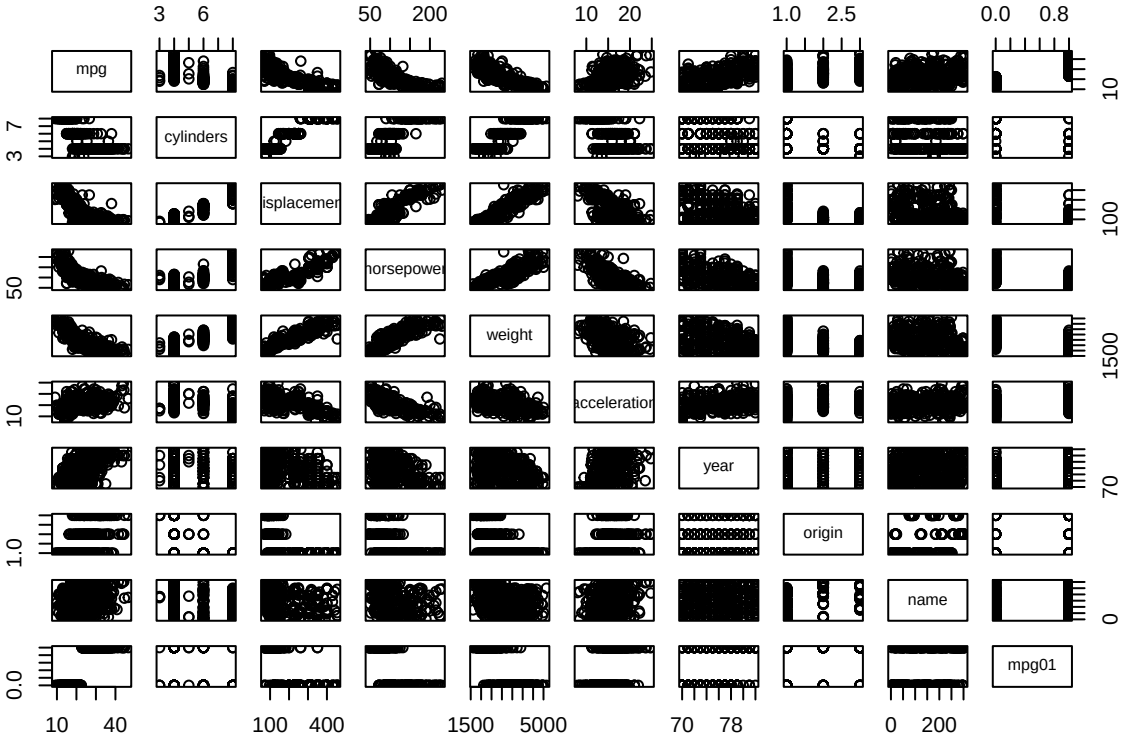
Explore the data graphically in order to investigate the association between mgp01 and the other features. Which of the other features seem most likely to be useful in predicting mgp01? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.

Answer:

```
#colnames(Auto)
```

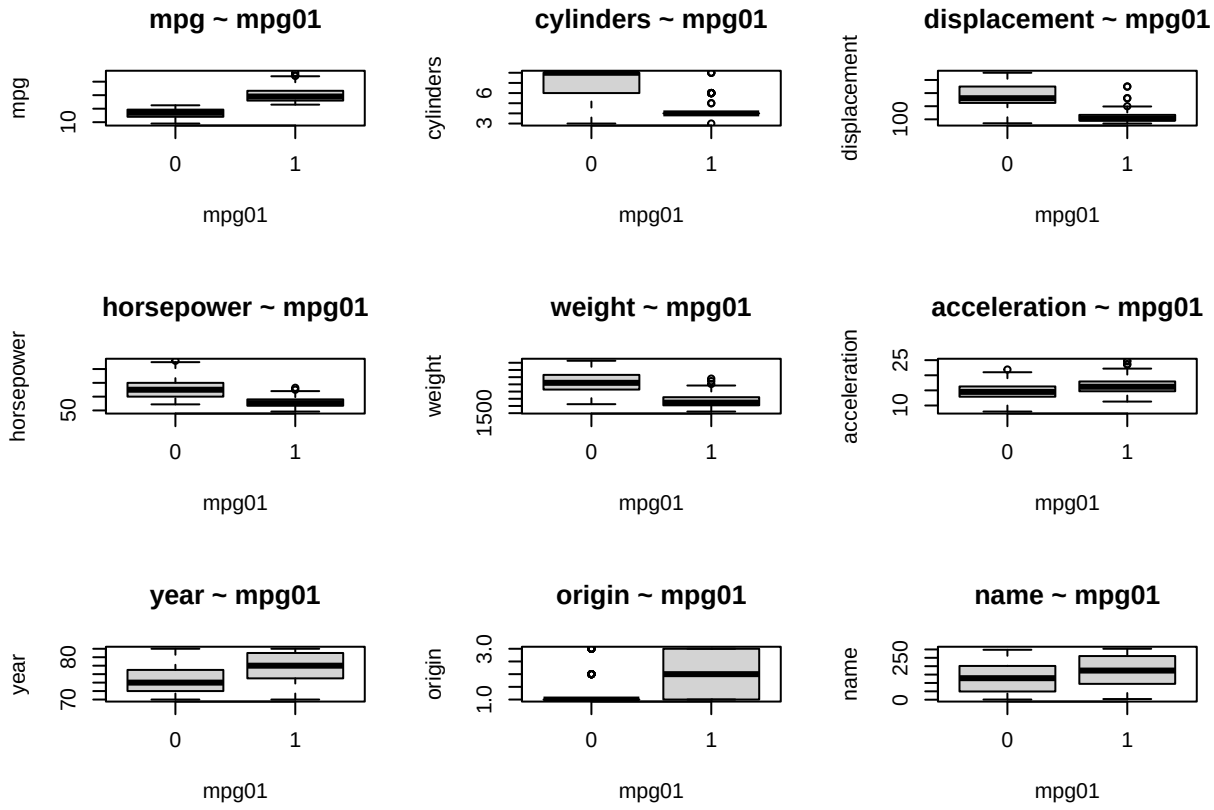
Scatter plot:

```
pairs(Auto)
```



box plot:

```
par(mfrow=c(3,3))
boxplot(mpg ~ mpg01, data = Auto, main = "mpg ~ mpg01")
boxplot(cylinders ~ mpg01, data = Auto, main = "cylinders ~ mpg01")
boxplot(displacement ~ mpg01, data = Auto, main = "displacement ~ mpg01")
boxplot(horsepower ~ mpg01, data = Auto, main = "horsepower ~ mpg01")
boxplot(weight ~ mpg01, data = Auto, main = "weight ~ mpg01")
boxplot(acceleration ~ mpg01, data = Auto, main = "acceleration ~ mpg01")
boxplot(year ~ mpg01, data = Auto, main = "year ~ mpg01")
boxplot(origin ~ mpg01, data = Auto, main = "origin ~ mpg01")
boxplot(name ~ mpg01, data = Auto, main = "name ~ mpg01")
```



In the scatterplot, we should choose values that have significantly different distributions when mpg01 is at 0 and 1. They are displacement, horsepower, weight, acceleration.

In the boxplot, we should choose the one with a relatively large difference between the two variable , for example, the difference between ourlier and the average value is relatively large. They are cylinders, displacement, horsepower, weight, year, origin.

Finally, these few variables that may affect mpg01 were selected are: cylinders, displacement, horsepower, weight.

Question 3

Split the data into a training set of size 300 and a test set of size 92

Answer:

```
# get the row number
dim(Auto)
```

```
## [1] 392  10
```

Total row is 392.

```
# randomly select 300 as training data
training_num <- sample(c(1:392),size=300)
training_num <- sort(training_num)
```

```
# Split the data into a training set of size 300 and a test set of size 92.
training_set <- Auto[training_num,]
test_set <- Auto[-training_num,]
dim(training_set)
```

```
## [1] 300  10
```

```
dim(test_set)
```

```
## [1] 92 10
```

I randomly select 300 as training data and the remaining 92 as testing data.

Question 4

Answer:

```
library(MASS)
# training
lda.fit <-  lda(mpg01 ~ cylinders + displacement + horsepower + weight, data = training_set)

# predict
lda.pred <-  predict(lda.fit, test_set)$class

cat("predicted result:\n")
```

```
## predicted result:
```

```
lda.pred
```

```
##  [1] 0 0 1 1 1 0 0 0 0 1 1 1 1 1 1 0 0 0 1 1 1 0 0 0 0 1 1 0 0 0 0 0 0 0 0 1 1 1
## [39] 1 1 0 0 0 1 0 1 1 0 0 0 0 1 1 1 1 0 0 0 0 0 0 1 1 1 1 1 0 1 0 1 1 1 1 0 1 1
## [77] 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 1
## Levels: 0 1
```

```
cat("test data:\n")
```

```
## test data:
```

```
test_set$mpg01
```

```
##  [1] 0 0 0 1 0 0 0 0 0 0 1 1 1 1 1 0 0 0 0 1 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 1 1
## [39] 1 0 0 0 0 0 0 1 1 0 0 0 0 1 1 1 1 0 0 0 0 0 0 1 1 0 0 1 0 1 1 1 1 1 1 0 1 1
## [77] 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1
```

```
# get the test error
error <- mean(lda.pred != test_set$mpg01)
cat(paste("\nThe test error of LDA model is :",error))
```

```
##
## The test error of LDA model is : 0.119565217391304
```

Note: Since 300 data are randomly selected, the results will be different each time. If you run my code, you may get different results.

Question 5

Perform QDA on the training data in order to predict mgp01 using the variables that seemed most associated with mpg01 in (b). What is the test error of the model obtained?

Answer:

```
# training
qda.fit <- qda(mpg01 ~ cylinders + displacement + horsepower + weight, data = training_set)

# predict
qda.pred <- predict(qda.fit, test_set)$class

cat("predicted result:\n")
```

```
## predicted result:
```

```
qda.pred
```

```
##  [1] 0 0 0 1 0 0 0 0 0 1 1 1 1 1 1 0 0 0 1 1 1 0 0 0 0 1 1 0 0 0 0 0 0 0 0 1 1 1
## [39] 1 1 0 0 0 0 0 1 1 0 0 0 0 1 1 1 1 0 0 0 0 0 0 1 1 1 1 1 0 1 0 1 1 1 1 0 1 1
## [77] 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 1
## Levels: 0 1
```

```
cat("test data:\n")
```

```
## test data:
```

```
test_set$mpg01
```

```
##  [1] 0 0 0 1 0 0 0 0 0 0 0 1 1 1 1 1 0 0 0 0 1 1 0 0 0 0 1 0 0 0 0 0 0 0 0 1 1 1
## [39] 1 0 0 0 0 0 0 0 1 1 0 0 0 0 1 1 1 1 0 0 0 0 0 0 1 1 0 0 1 0 1 1 1 1 1 1 0 1 1
## [77] 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1
```

```
# get the test error
error_qda <- mean(qda.pred != test_set$mpg01)
cat(paste("\nThe test error of QDA model is :",error_qda))
```

```
##
## The test error of QDA model is : 0.0869565217391304
```

Note: Since 300 data are randomly selected, the results will be different each time. If you run my code, you may get different results.