­

Ch2. Statistical Learning

-Lab & Exercises 9, 10

­­

**학번 : 182STG12**

**이름 : 오혜윤**

c

Problem Description

In Chapter 2 lab, we learn about the R functions regarding vector, matrix and graphics. For manipulating data first, using the basic functions right is important. There is a description about the function called persp() in this lab. Most people are not familiar with the function because they do not usually use 3-D plot for EDA. However, it is an useful function for drawing a map or showing the change of x and y.

In Chapter 2 exercises, using mpg and Boston housing dataset, we want to investigate the predictors. ‘mpg’ dataset is about cars and there are 11 variables in this dataset. Also, Boston housing dataset is about houses in Boston and there are 14 variables in this dataset. By creating some plots highlighting the relationships among the predictors and using some summary functions, we can understand what relationship exists between the variables and what characteristics are in the variables.

Results

**CH2. Lab Review**

In the lab, the first part was about simple R commands such as vector and matrix. I already knew them, so it was not that surprising or helpful. In the second part, there were some commands about graphics. I first knew the function called persp() through this lab. ‘persp()’ can be used to produce a three-dimensional plot. Also, the arguments theta and phi control the angles at which the plot is viewed. It was amazing and helpful. I did not know there was the function of making a 3D plot. I think I should try it next time.

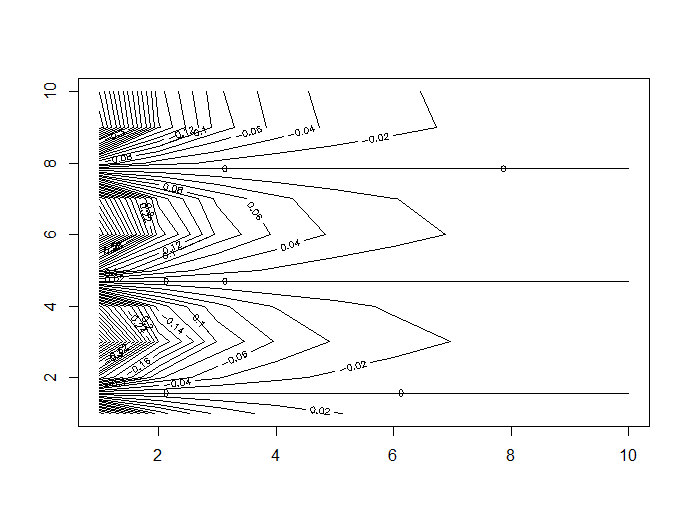


Figure 1: using contour()

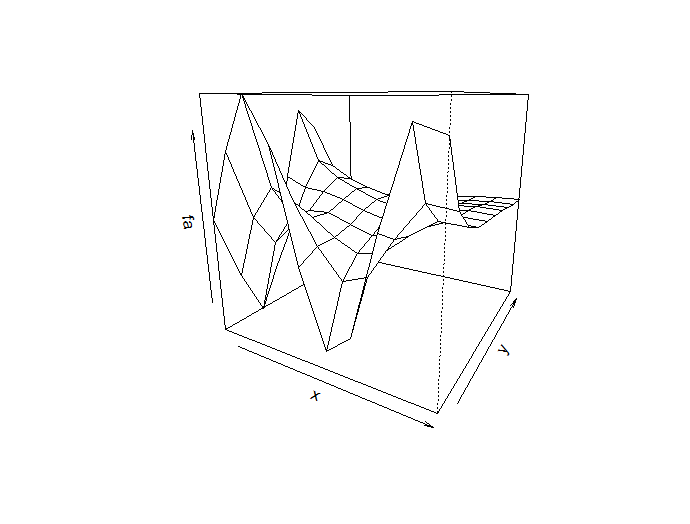


Figure 2 : using persp()

**CH2 Exercises**

**9. This exercise involves the Auto data set studied in the lab. Make sure that the missing values have been removed from the data.**

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

Quantitative means it can be counted. On the other hand, qualitative means it is a description. However, there are also qualitative variables that have been assigned a number. For example, in Auto data, a variable called “origin” means Origin of car (1. American, 2. European, 3. Japanese). Therefore, we can divide the predictors like this.

\* quantitative: mpg, cylinders, displacement, horsepower, weight, acceleration, year

\* qualitative: origin, name

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

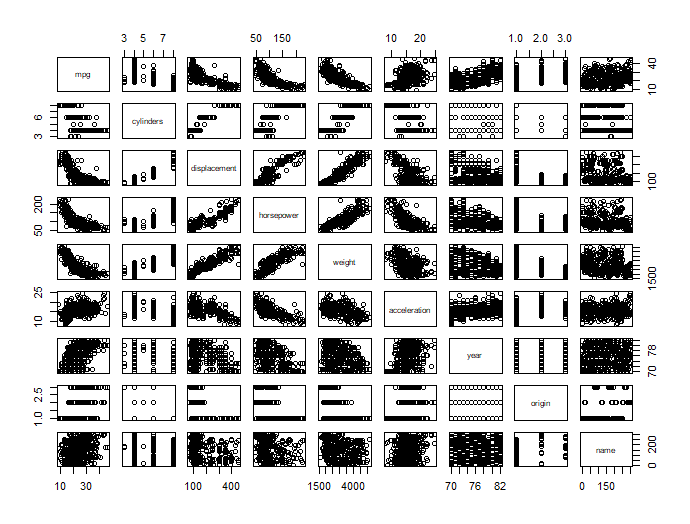
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Mpg | cylinders | displacement | horsepower |
| range | 9.0~46.6 | 3~8 | 68~455 | 46~230 |
|  | **Weight** | **acceleration** | **Year** |  |
| range | 1613~5140 | 8.0~24.8 | 70~82 |  |

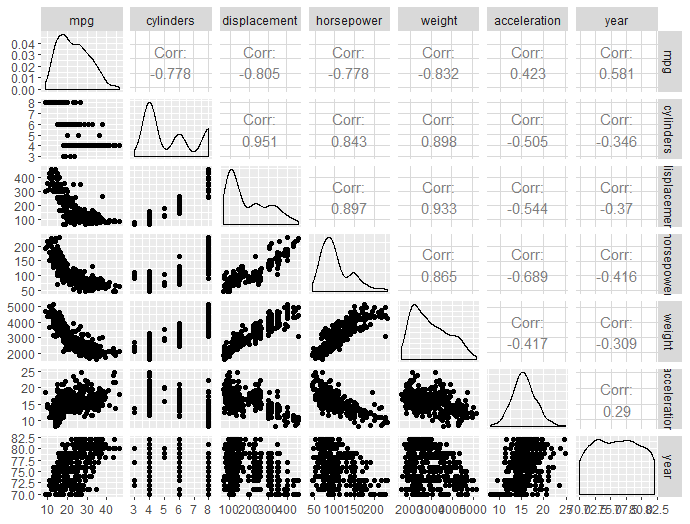
**(c) What is the mean and standard deviation of each quantitative predictor?**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | mpg | cylinders | displacement | horsepower |
| Mean | 23.45 | 5.472 | 194.4 | 104.5 |
| std | 7.8050 | 1.7058 | 104.644 | 38.4912 |
|  | **weight** | **acceleration** | **Year** |  |
| Mean | 2978 | 15.54 | 75.98 |  |
| std | 849.4026 | 2.7589 | 3.6837 |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | mpg | cylinders | displacement | horsepower |
| Range | 11.0~46.6 | 3~8 | 68~455 | 46~230 |
| Mean | 24.4044 | 5.3734 | 187.2405 | 100.7215 |
| Std | 7.8673 | 1.6542 | 35.7089 | 811.3002 |
|  | **weight** | **acceleration** | **Year** |  |
| Range | 1649~4997 | 8.5~24.8 | 70~82 |  |
| Mean | 2935.972 | 15.7269 | 77.1456 |  |
| std | 99.6784 | 2.6937 | 3.1062 |  |

**(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?**

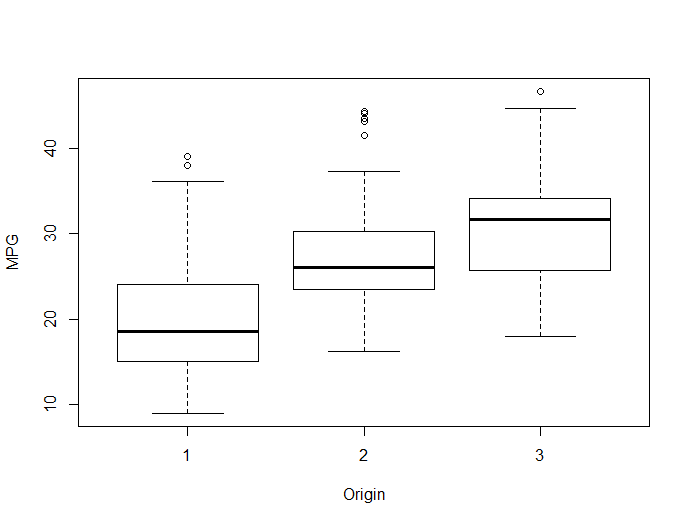
**(e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.**



\* `mpg` is negatively correlated with `cylinders`, `displacement`, `horsepower`, and `weight`

\* `horsepower` is negatively correlated with `weight`

\* `mpg` mostly increases for newer model years

\* ‘cylinders’ is negatively related to the ‘acceleration’ and ‘horsepower’.

\* ‘mpg’ is positively related to the ‘origin’.

**(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.**

At first glance, the scatterplot matrix of the quantitative variables gives us an idea

of the relationship between ‘mpg’ and the other quantitative variables.

\* mpg ~ cylinders, displacement, horsepower, weight: negative relationship

\* mpg ~ acceleration: positive relationship (seems unusual)

\* mpg ~ year: positive relationship

In addition, as previously mentioned, we observed a positive trend between mpg and the qualitative variable ‘origin’.

A multi-linear regression model with both quantitative and qualitative variables:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| cylinders | 1 | 14403.1 | 14403.1 | 1317.3788 | < 2.2e-16 \*\*\* |
| year | 1 | 2629.2 | 2629.2 | 240.4777 | < 2.2e-16 \*\*\* |
| weight | 1 | 2222.7 | 2222.7 | 203.3031 | < 2.2e-16 \*\*\* |
| horsepower | 1 | 1.6 | 1.6 | 0.1465 | 0.7021 |
| displacement | 1 | 10.8 | 10.8 | 0.9881 | 0.3208 |
| acceleration | 1 | 8.2 | 8.2 | 0.7539 | 0.3858 |
| origin | 2 | 356.0 | 178.0 | 16.2787 1. | 1.639e-07 \*\*\* |
| Residuals | 383 | 4187.4 | 10.9 | 1317.3788 |  |

Based on the anova table, we know that the predictors ‘horsepower, ’displacement’, and ‘acceleration’ are not statistically significant in estimating the mpg.

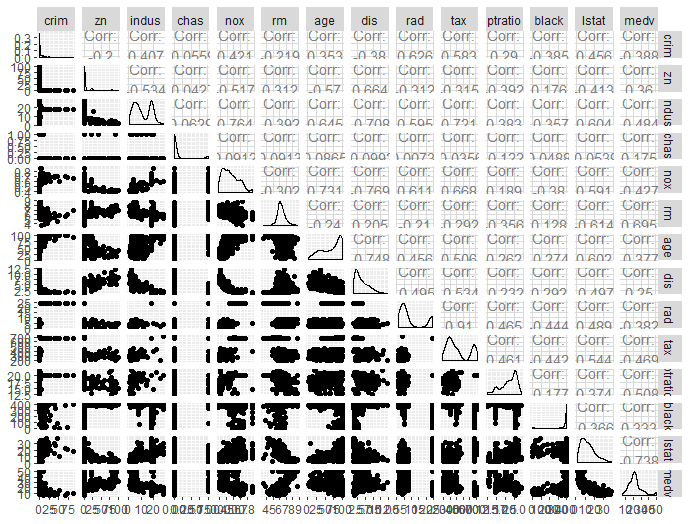
**10. This exercise involves the Boston housing data set.**

**(a) To begin, load in the Boston data set. The Boston data set is part of the MASS library in R. How many rows are in this data set? How many columns? What do the rows and columns represent?**

There are 506 rows and 14 columns in Boston data set.

Each row represents the set of observations for a given Neighborhood in Boston. Each column represents each predictor variable for which an observation was made in 506 neighborhoods of Boston.

**(b) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.**

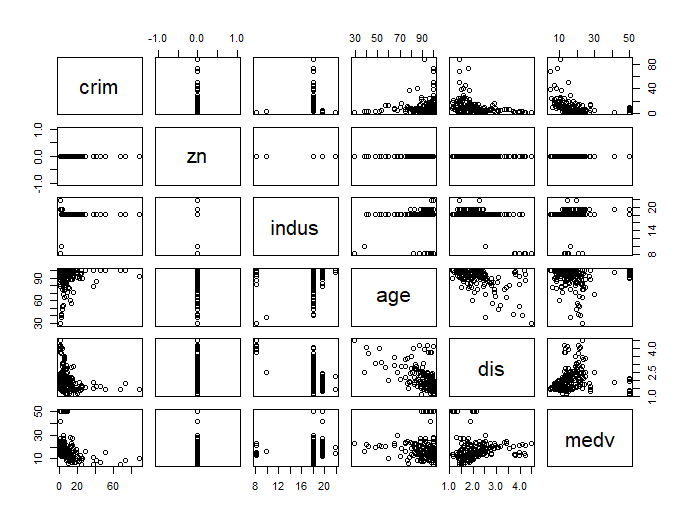
We can see high correlations between ‘medv’ & ‘lstat’ and also ‘medv’ & ‘rm’. ‘lstat’ is negatively correlated to ’medv’, whereas ‘rm’ is positively correlated which would be expected as the higher value areas would typically have larger dwellings. Lower correlation of median value is seen with ‘chas’ and ‘dis’. Also, there is a strong negative correlation between age of homes and distance to employment centers. The further homes are from an employment center, the newer they tend to be.

**(c) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.**

Of the four predictors with the strongest correlations with per capita crime rate, accessibility to radial highways (rad) and full-value property-tax rate (tax) have moderately strong positive relationships with crime rate.

|  |  |
| --- | --- |
| corr | crime |
| Zn | -0.2005 |
| indus | 0.4066 |
| chas | -0.0559 |
| Nox | 0.4210 |
| Rm | -0.2193 |
| Age | 0.3527 |
| Dis | -0.3797 |
| Rad | 0.6255 |
| Tax | 0.5828 |
| ptratio | 0.2899 |
| black | -0.3851 |
| lstat | 0.4556 |
| medv | -0.3883 |

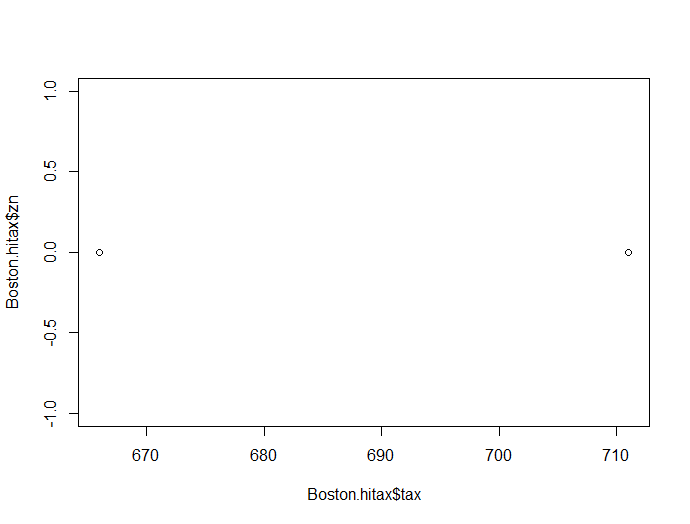
Many areas have low crime (per capita crime rate <20), regardless of the accessibility to highways, tax rate or lower status of the population. However, there are a few outlying suburbs with very high crime rates.

 **(d) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.**

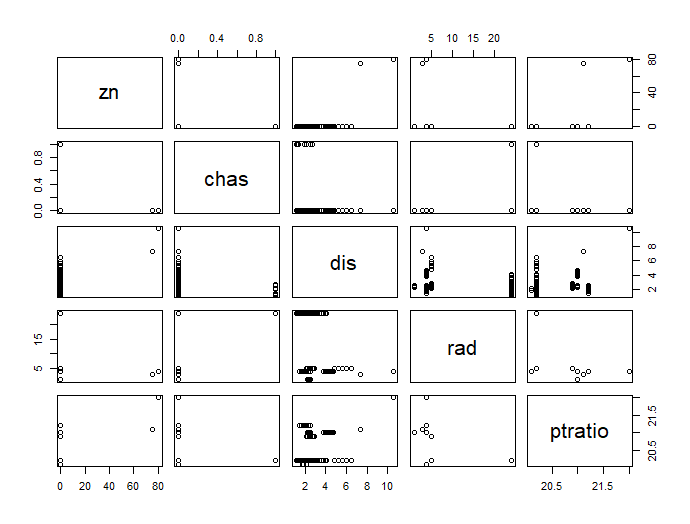
1. Of all observations, there were approximately 66% with lower than 1% crime rate per capita. The range for crim is (0.006, 88.976). The suburbs that appear to have high crime rates (> 1%) generally have:

* No residential land zoned for lots over 25,000 sq ft
* Less than 18% of non-retail business acres per town
* A higher proportion of older housing
* Tend to be closer to employment centers
* Lower median home values

2. The range of tax rates appear to form two clusters – 187 to 469 and 666 to 711. No residential land zoned for lots over 25,000 sq ft belong in the higher tax group.



3. The range for ptratio is (12.6, 22.0). Of the observations, there are approximately 40% having a pupil-teacher ratio of 20-22. The group with a higher ptratio generally:

* Have very little residential land zoned for lots over 25,000 sq ft
* Are not along the Charles River
* Are closer to employment centers and radial highways

**(e) How many of the suburbs in this data set bound the Charles river?**

There are 35 suburbs in this data set.

**(f) What is the median pupil-teacher ratio among the towns in this data set?**

The median pupil-teacher ratio is 19.05.

**(g) Which suburb of Boston has lowest median value of owner occupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your findings.**

The crime rate is on the high side for the two suburbs, as is indus, tax, ptratio, black, and lstat. The age and rad variables are equal to the max values. These two observations are very similar across all predictors.

**(h) In this data set, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.**

64 suburbs average more than seven rooms per dwelling. Also, 13 suburbs average more than eight rooms per dwelling.

Considering the small sample size of 13 observations for suburbs averaging more than eight rooms, the ranges are very similar to those of the entire dataset. The one major exception is the black predictor, which is near the top of the range in all observations in the subset.

river?

Discussion

After reviewing the datasets and using graphical analysis, we found many relationships between the variables. In the mpg dataset, mpg has negative relationship with many variables such as cylinders, displacement, horsepower, weight. Also, mpg has positive relationship with acceleration and it was unusual. On the other hand, in the Boston housing dataset, we can see high correlations between ‘medv’ & ‘lstat’ and also ‘medv’ & ‘rm’. ‘lstat’ is negatively correlated to ’medv’, whereas ‘rm’ is positively correlated which would be expected as the higher value areas would typically have larger dwellings.

By investigating the datasets, we can find many interesting things. Especially, by creating correlation matrix and graphs, we can understand datasets better. The function called ggpairs() was useful for showing the relationships between the variables entirely. We expect to use that function for analyzing data next time.

Appendix (R

**R codes**

#ex9----------------------------------------------------------

install.packages("ISLR")

require(ISLR)

data(Auto)

fix(Auto)

Auto = na.omit(Auto)

Auto$origin <- as.factor(Auto$origin)

#b-----------------------------------------

quant\_data=Auto[,c(1:7)]

apply(quant\_data,2, range)

#c-----------------------------------------

colMeans(quant\_data)

apply(quant\_data,2, sd)

#d-----------------------------------------

Auto2 = Auto[-c(10:85),]

quant\_data2=Auto2[,c(1:7)]

apply(quant\_data2,2, range)

colMeans(quant\_data2)

apply(quant\_data,2, sd)

#e-----------------------------------------

library(GGally)

library(ggplot2)

ggpairs(quant\_data)

plot( Auto$mpg ~ Auto$origin , xlab ="Origin", ylab="MPG")

#f------------------------------------------

model <- lm(mpg ~ cylinders+year+weight+horsepower+displacement+acceleration+origin,

data= Auto)

t(coefficients(model))

anova(model )

#ex10-----------------------------------------------------------------------

#a----------------------------------------

library(MASS)

head(Boston)

dim(Boston)

#b-----------------------------------------

Boston$chas <- as.numeric(Boston$chas)

Boston$rad <- as.numeric(Boston$rad)

pairs(Boston)

ggpairs(Boston)

#c----------------------------------------

cor(Boston)

#d----------------------------------------

Boston.hicrim <- subset(Boston, crim > 1)

Boston.hitax <- subset(Boston, tax > 500)

Boston.hiptratio <- subset(Boston, ptratio > 20)

pairs(Boston.hicrim[,c(1,2,3,7,8,14)])

plot(Boston.hitax$tax, Boston.hitax$zn)

pairs(Boston.hiptratio[,c(2,4,8,9,11)])

#e----------------------------------------

length(Boston$chas[Boston$chas==1])

#f----------------------------------------

median(Boston$ptratio)

#g------------------------------------------

g10 <- rbind(Boston[Boston$medv==min(Boston$medv),],sapply(Boston, range))

rownames(g10) <- c("Lowest medv 1", "Lowest medv 2", "Min", "Max")

g10

#h----------------------------------------

length(Boston$rm[Boston$rm>7])

length(Boston$rm[Boston$rm>8])

h10 <- rbind(sapply(Boston[Boston$rm>8,], range), sapply(Boston, range))

rownames(h10) <- c("Min rm>8", "Max rm>8", "Boston min", "Boston max")

h10