Jingchun Ma

contact:13470463996 | email:tunna_M@163.com (mailto:tunna_M@163.com) | Address:Shanghai

Education

Sept 2020 to current *School of Statistics and Management, Shanghai University of Finance and Economics*

- **Major**: Machine Learning, Database, Data Structure, Linear Model, Mathematical Analysis, Advanced Algebra

Interests

- **code**

  got excellent grades in programming courses
- **photography**

  tens of thousands of photos on the phone
- **swimming**

  always go for a swim

Future plan

1. Continue to study data science abroad
2. Hope to work in the Internet industry

Expect to learn from this class

1. Learn a lot about machine learning methods
2. Good command of R language
3. Improve academic writing skills

Exercise 2

1. **create the vector 1,1,1,1,1,2,2,2,2,2 with only rep() and name it x1.**

```
x1 = rep(c(1,2),each = 5)
x1
```

```
##  [1] 1 1 1 1 1 2 2 2 2 2
```

2. **create the vector 1,2,1,2,1,2,1,2,1,2 with only rep() and name it x2.**

```
x2 = rep(c(1,2),times = 5)
x2
```

```
##  [1] 1 2 1 2 1 2 1 2 1 2
```

3. **combine x1 and x2 into a matrix x.col by columns, i.e., x1 and x2 are the two columns of x. Hint: use cbind().**

```
x3 = cbind(x1,x2)
x3
```

```
##       x1 x2
##  [1,]  1  1
##  [2,]  1  2
##  [3,]  1  1
##  [4,]  1  2
##  [5,]  1  1
##  [6,]  2  2
##  [7,]  2  1
##  [8,]  2  2
##  [9,]  2  1
## [10,]  2  2
```

4. **combine x1 and x2 into a matrix x.row by rows, i.e., x1 and x2 are the two rows of x. Hint: use rbind().**

```
x4 = rbind(x1,x2)
x4
```

```
##    [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## x1    1    1    1    1    1    2    2    2    2     2
## x2    1    2    1    2    1    2    1    2    1     2
```

5. **find two ways to calculate the sum of each column of x.row. Hint: use apply().**

Method 1

```
apply(x4 , 2 , sum)
```

```
##  [1] 2 3 2 3 2 4 3 4 3 4
```

Method 2

```
colSums(x4)
```

```
##  [1] 2 3 2 3 2 4 3 4 3 4
```

Exercise 3

1. **How many rows are in this data set? How many columns? What do the rows and columns represent?**

```
library(ISLR2)
nrow(Boston)
```

```
## [1] 506
```

```
ncol(Boston)
```

```
## [1] 13
```

There are 506 rows and 13 colomns in the data set. The rows represent the total amount of data. The colomns represent the indicator.

2. **Which of the predictors are quantitative, and which are qualitative?**

```
names(Boston)
```

```
##  [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"
##  [8] "dis"     "rad"     "tax"     "ptratio" "lstat"   "medv"
```

```
summary(Boston)
```

```
##       crim                zn             indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
##       nox               rm             age              dis        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
##       rad              tax           ptratio          lstat      
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   : 1.73  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.: 6.95  
##  Median : 5.000   Median :330.0   Median :19.05   Median :11.36  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :12.65  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:16.95  
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :37.97  
##       medv      
##  Min.   : 5.00  
##  1st Qu.:17.02  
##  Median :21.20  
##  Mean   :22.53  
##  3rd Qu.:25.00  
##  Max.   :50.00  
```

chas and rad are qualitative.

The rest are quantitative data

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

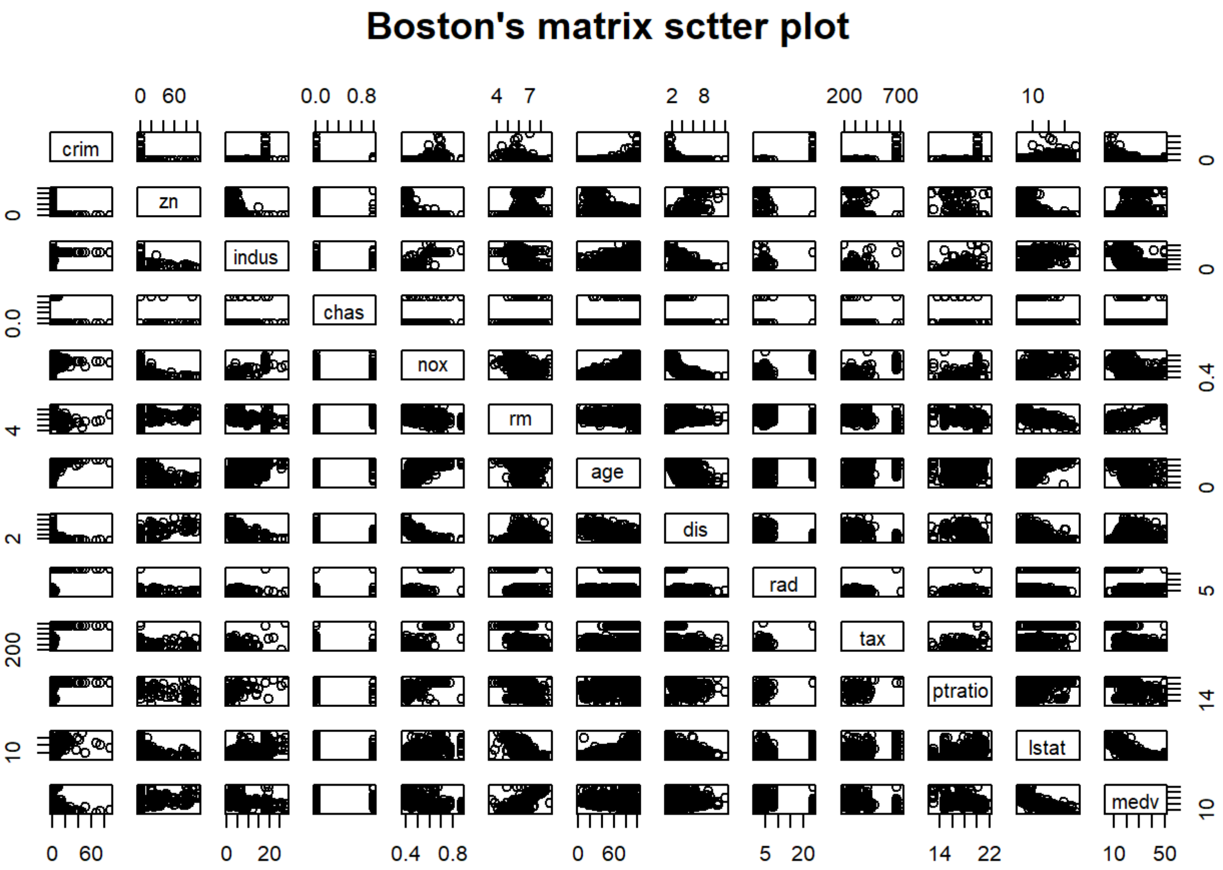
```
Boston1 <- Boston[,-c(4,9)]
len = matrix(0,11,2)

for (l in 1:11){
  len[l,]=range(Boston1[,l])
}
#提取变量名
name = matrix(names(Boston1[,1:11]),11,1)
len = cbind(name,len)
len=data.frame(len)
names(len)=c("变量名","最小值","最大值")
len
```

```
##      变量名   最小值   最大值
## 1      crim 0.00632 88.9762
## 2        zn       0     100
## 3     indus    0.46   27.74
## 4       nox   0.385   0.871
## 5        rm   3.561    8.78
## 6       age     2.9     100
## 7       dis  1.1296 12.1265
## 8       tax     187     711
## 9   ptratio    12.6      22
## 10    lstat    1.73   37.97
## 11     medv       5      50
```

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

```
pairs(Boston[,1:13],main="Boston's matrix sctter plot")
```



nox and dix are linearly and negatively correlated

rm and lstat are negatively correlated, but rm and medv are positively correlated.

lstat and medv are positively correlated.

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

Based on the picture from the previous question, crim and zn,crim and indus,crim and chas are linearly dependent. But there was no positive correlation or negetive correlation.That means these variables don’t change with crim.

6. **What is the mean and standard deviation of each quantitative predictor?**

```
len2=matrix(0,11,2)
for (l in 1:11){
  len2[l,1]=mean(Boston1[,l])#变量均值
  len2[l,2]=sd(Boston1[,l])#变量标准差

}
len2 = cbind(name,len2)
len2=data.frame(len2)
names(len2)=c("变量名","均值","标准差")
len2
```

```
##     变量名               均值             标准差
## 1     crim  3.61352355731225  8.60154510533249
## 2       zn  11.3636363636364  23.3224529945151
## 3    indus  11.1367786561265  6.86035294089759
## 4      nox 0.554695059288538 0.115877675667556
## 5       rm  6.28463438735178 0.702617143415323
## 6      age  68.5749011857708  28.1488614069036
## 7      dis  3.79504268774704  2.10571012662761
## 8      tax  408.237154150198  168.537116054959
## 9  ptratio  18.4555335968379  2.16494552371444
## 10   lstat  12.6530632411067  7.14106151134857
## 11    medv  22.5328063241107  9.19710408737982
```

7. **How many of the census tracts in this data set bound the Charles river?**

```
sum(Boston["chas"])
```

```
## [1] 35
```

There are 35 census tracts in this data set bound the Charles river

8. **What is the median pupil-teacher ratio among the towns in this data set?**

```
ptratio <- as.matrix(Boston["ptratio"])
median(ptratio)
```

```
## [1] 19.05
```

9. **Which census tract of Boston has lowest median value of owner-occupied homes? What are the values of the other predictors for that census tract, and how do those values compare to the**

overall ranges for those predictors? Comment on your findings.

```
age <- as.matrix(Boston["age"])
x <- which.min(age)
Boston[x,]
```

```
##       crim zn indus chas   nox   rm age    dis rad tax ptratio lstat medv
## 42 0.12744  0  6.91    0 0.448 6.77 2.9 5.7209   3 233    17.9  4.84 26.6
```

The No.42 census tract of Boston has lowest median value of owner-occupied homes. These values are small compared to the other rows.

10. **In this data set, how many of the census tracts average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the census tracts that average more than eight rooms per dwelling.**

```
rm7 <- nrow(Boston[Boston$rm > 7, ])
rm7
```

```
## [1] 64
```

```
rm8 <- nrow(Boston[Boston$rm > 8, ])
rm8
```

```
## [1] 13
```

```
print(Boston[Boston$rm > 8, ])
```

```
##        crim zn indus chas    nox    rm  age    dis rad tax ptratio lstat medv
## 98  0.12083  0  2.89    0 0.4450 8.069 76.0 3.4952   2 276    18.0  4.21 38.7
## 164 1.51902  0 19.58    1 0.6050 8.375 93.9 2.1620   5 403    14.7  3.32 50.0
## 205 0.02009 95  2.68    0 0.4161 8.034 31.9 5.1180   4 224    14.7  2.88 50.0
## 225 0.31533  0  6.20    0 0.5040 8.266 78.3 2.8944   8 307    17.4  4.14 44.8
## 226 0.52693  0  6.20    0 0.5040 8.725 83.0 2.8944   8 307    17.4  4.63 50.0
## 227 0.38214  0  6.20    0 0.5040 8.040 86.5 3.2157   8 307    17.4  3.13 37.6
## 233 0.57529  0  6.20    0 0.5070 8.337 73.3 3.8384   8 307    17.4  2.47 41.7
## 234 0.33147  0  6.20    0 0.5070 8.247 70.4 3.6519   8 307    17.4  3.95 48.3
## 254 0.36894 22  5.86    0 0.4310 8.259  8.4 8.9067   7 330    19.1  3.54 42.8
## 258 0.61154 20  3.97    0 0.6470 8.704 86.9 1.8010   5 264    13.0  5.12 50.0
## 263 0.52014 20  3.97    0 0.6470 8.398 91.5 2.2885   5 264    13.0  5.91 48.8
## 268 0.57834 20  3.97    0 0.5750 8.297 67.0 2.4216   5 264    13.0  7.44 50.0
## 365 3.47428  0 18.10    1 0.7180 8.780 82.9 1.9047  24 666    20.2  5.29 21.9
```

Notably, these areas are much closer to the five centers of Boston. There are fewer people of lower status, and home ownership is also more expensive.