Exam Chapter 2

Wenduo Wang

July 19, 2016

Problem 10

(a, b) Load the MASS library and explore the Boston dataset a little bit.

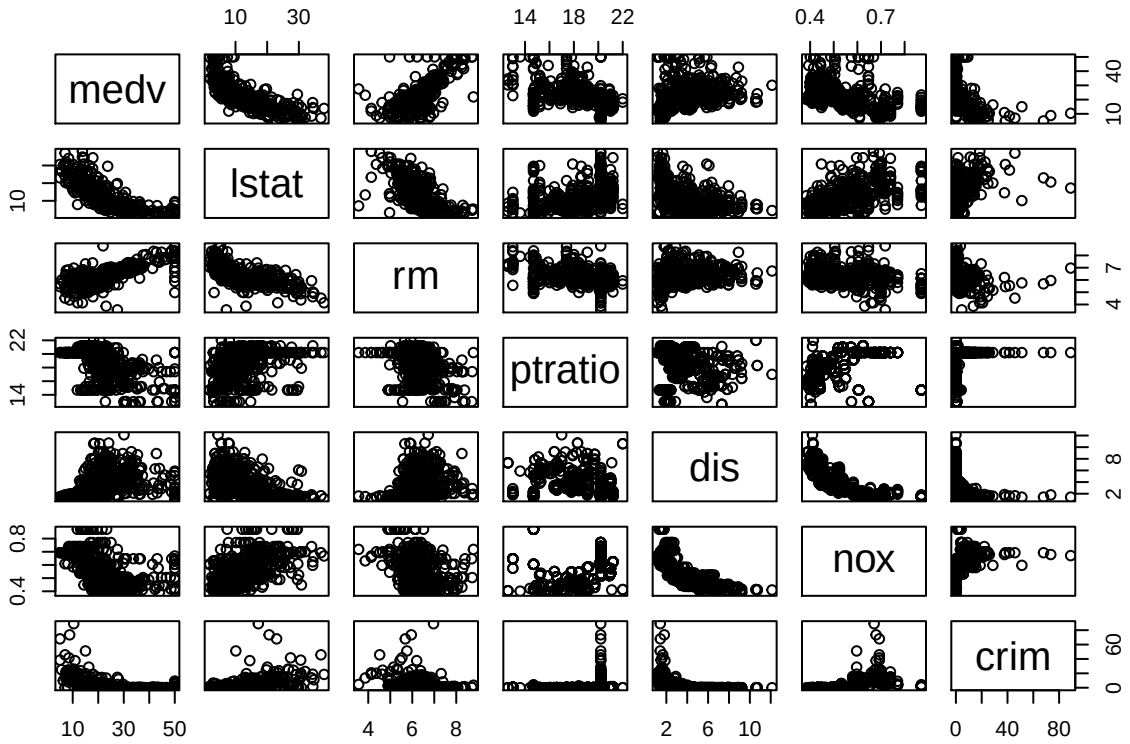
```
library(MASS)
summary(Boston)
```

```
##       crim                zn              indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08204   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          black       
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23  
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90  
##      lstat            medv      
##  Min.   : 1.73   Min.   : 5.00  
##  1st Qu.: 6.95   1st Qu.:17.02  
##  Median :11.36   Median :21.20  
##  Mean   :12.65   Mean   :22.53  
##  3rd Qu.:16.95   3rd Qu.:25.00  
##  Max.   :37.97   Max.   :50.00
```

```
cat("There are", nrow(Boston), "rows and", ncol(Boston), "columns in the Boston dataset")
```

```
## There are 506 rows and 14 columns in the Boston dataset
```

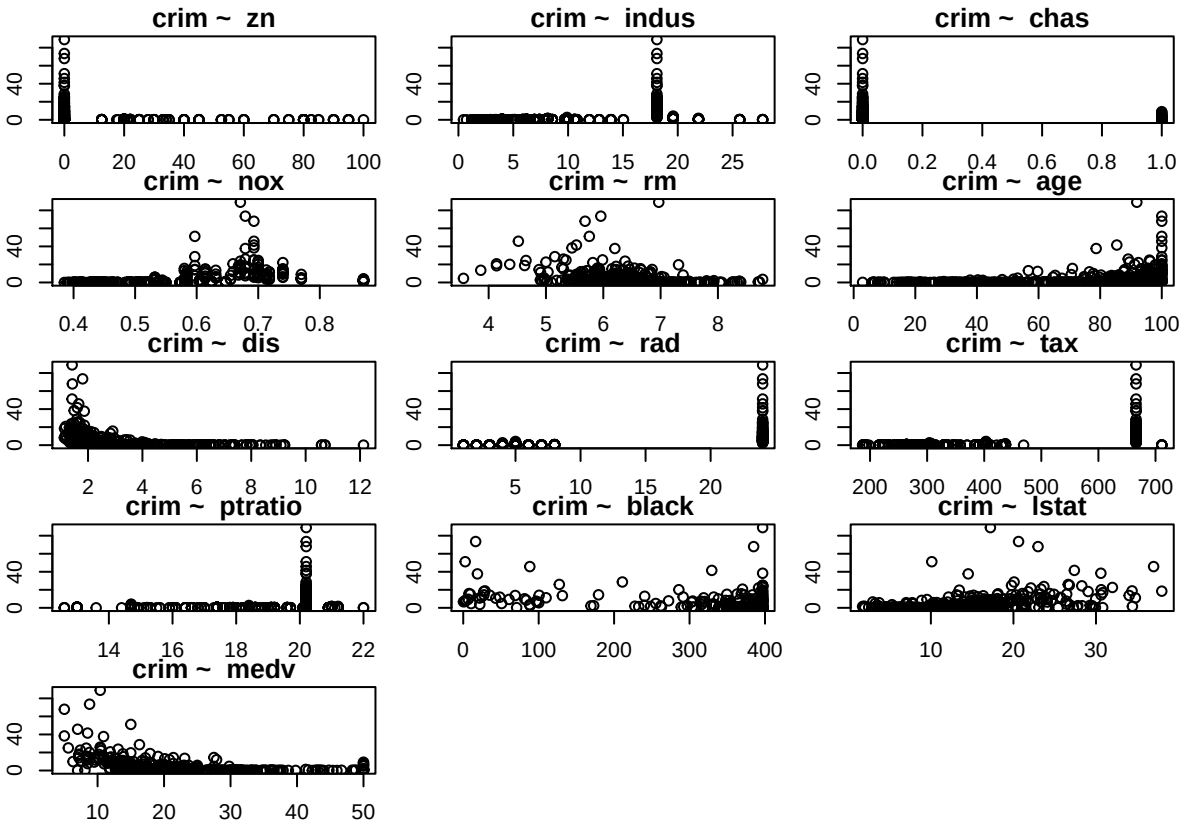
```
pairs(medv~lstat+rm+ptratio+dis+nox+crim, data=Boston)
```



From the scatterplot it is observed that the median value of owner-occupied homes (`medv`) has a strong negative correlation with low status population and strong positive correlation with room numbers (`rm`). However, such correlation is less obvious for other variables.

(c) Let's explore the correlation between per capita crime rate by town (`crim`) against other variables.

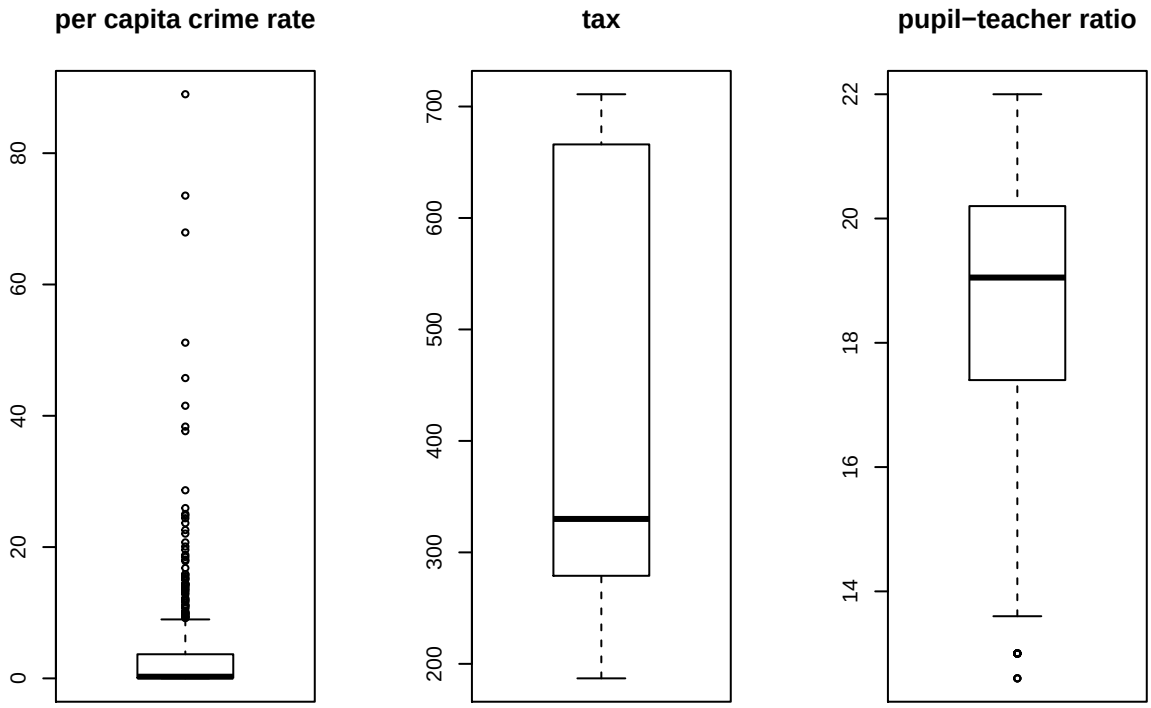
```
par(mfrow=c(5,3), mar=rep(1.5, 4), oma=rep(1,4))
for (i in c(1:ncol(Boston))){
    if (colnames(Boston)[i]=="crim"){
        next
    }
    plot(y=Boston$crim, x=Boston[,i], ylab="", xlab="", main=paste("crim ~ ", colnames(Boston)[i]))
}
```



From the output plot it is difficult to assert correlations against per capita crime rate. However, there is a negative correlation between `crim` and `dis`, which also applies to `medv`, which means the per capita crime rate tends to be higher in areas closer to the five Boston employment centres (or where the median house value is lower). In contrast, in places with higher `lstat`, lower status population, `crim` is also higher, showing a positive correlation. Similar relationship is also implied between `crim` and `nox`, nitrogen oxides concentration.

(d)

```
library(car)
par(mfrow=c(1,3))
boxplot(Boston$crim, main="per capita crime rate", id.n=Inf)
boxplot(Boston$tax, main="tax", id.n=Inf)
boxplot(Boston$ptratio, main="pupil-teacher ratio", id.n=Inf)
```




```
cat("The stats of per capita crime rate is below:")
```

```
## The stats of per capita crime rate is below:
```

```
summary(Boston$crim)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##   0.00632   0.08204   0.25650   3.61400   3.67700 88.98000
```

```
cat("The stats of full-value property-tax rate per $10,000 is below:")
```

```
## The stats of full-value property-tax rate per $10,000 is below:
```

```
summary(Boston$tax)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   187.0   279.0   330.0   408.2   666.0   711.0
```

```
cat("The stats of pupil-teacher ratio is below:")
```

```
## The stats of pupil-teacher ratio is below:
```

```
summary(Boston$ptratio)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.60   17.40   19.05   18.46   20.20   22.00
```

The box plot shows that several areas have extraordinarily high per capita crime rate, and a couple are particularly low in pupil-teacher ratio. Yet the property tax rate is within the normal range among all areas.

(e)

```
cat("There are", sum(Boston$chas==1), "suburbs bound the Charles River in the dataset.")
```

```
## There are 35 suburbs bound the Charles River in the dataset.
```

(f)

```r
cat("The median value of the pupil-teacher ratio is", median(Boston$ptratio), "in the Boston dataset.")
```

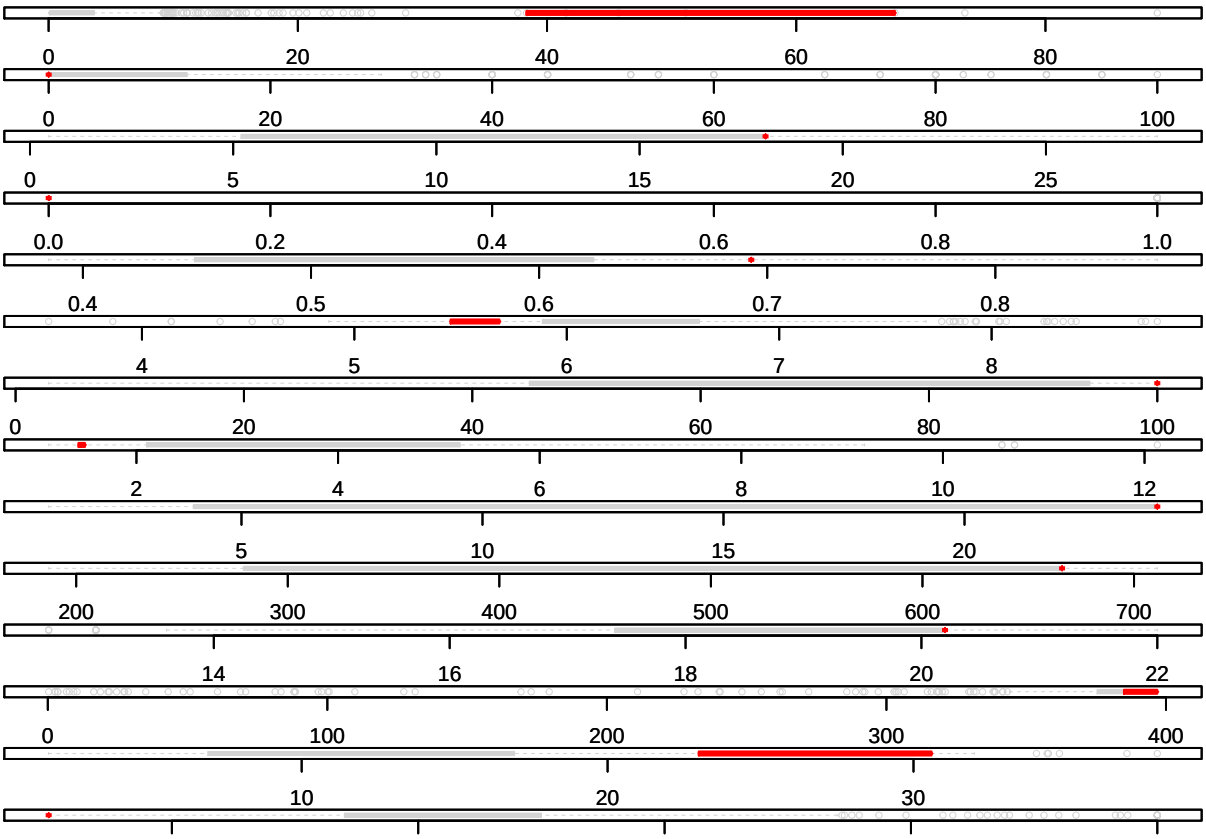
```
## The median value of the pupil-teacher ratio is 19.05 in the Boston dataset.
```

(g)

```r
library(reshape2)
low_medv <- Boston[Boston$medv==min(Boston$medv),]
if (nrow(low_medv)>1){
    cat("There are", nrow(low_medv), "suburbs with the lowest median house value.")
    cat("Their row numbers are", paste(which(Boston$medv==min(Boston$medv))), ".")
} else {
    cat("Suburb", which(Boston$medv==min(Boston$medv)), "has the lowest median house value.")
}
```

```
## There are 2 suburbs with the lowest median house value.Their row numbers are 399 406 .
```

```r
par(mfrow=c(dim(Boston)[2], 1), mar=rep(1, 4))
for (i in c(1:dim(Boston)[2])){
    boxplot(Boston[i], border="lightgray", col="lightgray", horizontal=TRUE, lwd=0.5)
    boxplot(low_medv[i], border="red", col="red", horizontal=TRUE, lwd=1, add=TRUE)
}
```



To compare the predictors of low median house value suburbs with the sample population, the boxplots of all predictors are put together. From top to bottom are the `crim` to `medv` The lightgray areas and points represents the sample population, while red areas and dark lines are the low median house value suburbs. The pattern demonstrates that suburbs whose median value of houses also have predictors away from the sample mean/median. For example, on the top of the boxplots is the per capita crime rate predictor, the low median house value suburbs (indicated by the red shade) have significantly higher values than the sample population (rendered in lightgray). After analysing each predictor, it is found that the low median house

value is positively related to `crim`, `indus`, `nox`, `age`, `rad`, `ptratio`, `tax` and `lstat`; while `zn`, `chas`, `rm` and `dis` is negatively related to the low median house value.

(h)

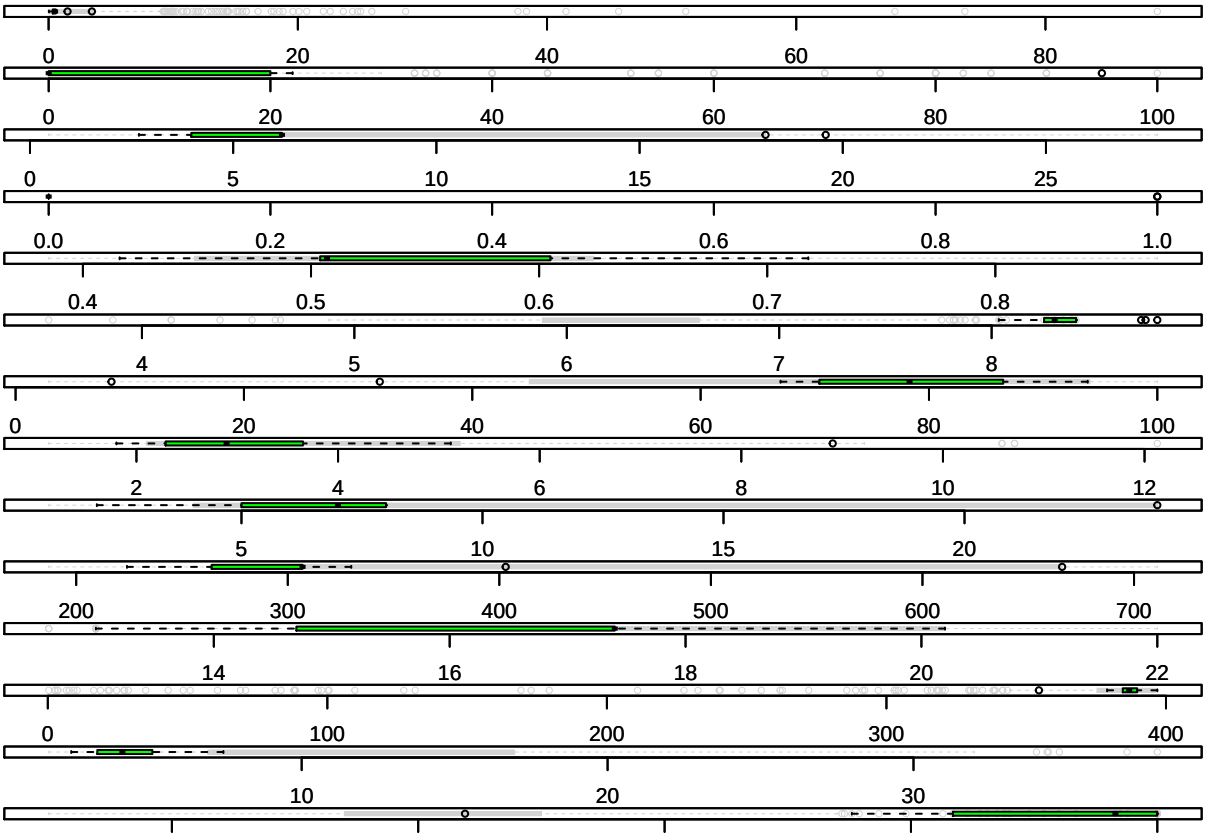
```
cat(sum(Boston$rm > 7), "suburbs average more than seven rooms per dwelling.")
```

```
## 64 suburbs average more than seven rooms per dwelling.
```

```
cat(sum(Boston$rm > 8), "suburbs average more than eight rooms per dwelling.")
```

```
## 13 suburbs average more than eight rooms per dwelling.
```

```
more_room <- Boston[Boston$rm > 8,]
par(mfrow=c(dim(Boston)[2], 1), mar=rep(1, 4))
for (i in c(1:dim(Boston)[2])){
    boxplot(Boston[i], border="lightgray", col="lightgray", horizontal=TRUE, lwd=0.5)
    boxplot(more_room[i], borde="green", col="green", horizontal=TRUE, lwd=1, add=TRUE)
}
```



Similar to the analysis of low median house value suburbs, the boxplots of suburbs averaging more than 8 rooms per dwelling is plotted (green) against the sample population (lightgray). Closer observation reveals that these suburbs, relative to the sample, also have considerably lower `crim`, `indus`, `rad`, `tax`, `lstat` and `medv`.