3.1 Boston数据

徐平峰

长春工业大学

2021/3/27

Boston房价数据

```
setwd("E:/teaching_plan_notes/msa11091083/rmd")#设置工作目录
getwd()#获得工作目录
```

```
# [1] "E:/teaching_plan_notes/msa11091083/rmd"
```

```
w=read.csv("MVAPureData/BostonHousing2.csv")#读取Boston房价数据
dim(w)
```

```
# [1] 506  19
```

```
#View(w)
names(w)
```

```
#  [1] "town"    "tract"   "lon"     "lat"     "medv"    "cmedv"   "crim"
#  [8] "zn"      "indus"   "chas"    "nox"     "rm"      "age"     "dis"
# [15] "rad"     "tax"     "ptratio" "b"       "lstat"
```

```
w=w[,-c(1:5)]#去掉前5列变量，只分析后14个变量
dim(w)
```

```
# [1] 506  14
```

Boston房价数据

```
head(w)#显示数据的前6行
```

```
#   cmedv    crim zn indus chas   nox    rm  age    dis rad tax ptratio
# 1  24.0 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.9
# 2  21.6 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.9
# 3  34.7 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.8
# 4  33.4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.6
# 5  36.2 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.9
# 6  28.7 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.1
#   lstat
# 1  4.98
# 2  9.14
# 3  4.03
# 4  2.94
# 5  5.33
# 6  5.21
```

Boston房价数据

```
tail(w)#显示数据的后6行
```

```
#     cmedv    crim zn indus chas   nox    rm  age    dis rad tax ptratio
# 501  16.8 0.22438  0  9.69    0 0.585 6.027 79.7 2.4982   6 391    19.2 396
# 502  22.4 0.06263  0 11.93    0 0.573 6.593 69.1 2.4786   1 273    21.0 391
# 503  20.6 0.04527  0 11.93    0 0.573 6.120 76.7 2.2875   1 273    21.0 396
# 504  23.9 0.06076  0 11.93    0 0.573 6.976 91.0 2.1675   1 273    21.0 396
# 505  22.0 0.10959  0 11.93    0 0.573 6.794 89.3 2.3889   1 273    21.0 393
# 506  19.0 0.04741  0 11.93    0 0.573 6.030 80.8 2.5050   1 273    21.0 396
#     lstat
# 501 14.33
# 502  9.67
# 503  9.08
# 504  5.64
# 505  6.48
# 506  7.88
```

Boston房价数据

```
w$cmedv#w数据的变量名为cmedv的列
```

```
#   [1] 24.0 21.6 34.7 33.4 36.2 28.7 22.9 22.1 16.5 18.9 15.0 18.9 21.7 20.4
#  [16] 19.9 23.1 17.5 20.2 18.2 13.6 19.6 15.2 14.5 15.6 13.9 16.6 14.8 18.4
#  [31] 12.7 14.5 13.2 13.1 13.5 18.9 20.0 21.0 24.2 30.8 34.9 26.6 25.3 24.7
#  [46] 19.3 20.0 16.6 14.4 19.4 19.7 20.5 25.0 23.4 18.9 35.4 24.7 31.6 23.3
#  [61] 18.7 16.0 22.2 25.0 33.0 23.5 19.4 22.0 17.4 20.9 24.2 21.7 22.8 23.4
#  [76] 21.4 20.0 20.8 21.2 20.3 28.0 23.9 24.8 22.9 23.9 26.6 22.5 22.2 23.6
#  [91] 22.6 22.0 22.9 25.0 20.6 28.4 21.4 38.7 43.8 33.2 27.5 26.5 18.6 19.3
# [106] 19.5 19.5 20.4 19.8 19.4 21.7 22.8 18.8 18.7 18.5 18.3 21.2 19.2 20.4
# [121] 22.0 20.3 20.5 17.3 18.8 21.4 15.7 16.2 18.0 14.3 19.2 19.6 23.0 18.4
# [136] 18.1 17.4 17.1 13.3 17.8 14.0 14.4 13.4 15.6 11.8 13.8 15.6 14.6 17.8
# [151] 21.5 19.6 15.3 19.4 17.0 15.6 13.1 41.3 24.3 23.3 27.0 50.0 50.0 50.0
# [166] 25.0 50.0 23.8 23.8 22.3 17.4 19.1 23.1 23.6 22.6 29.4 23.2 24.6 29.9
# [181] 39.8 36.2 37.9 32.5 26.4 29.6 50.0 32.0 29.8 34.9 33.0 30.5 36.4 31.1
# [196] 50.0 33.3 30.3 34.6 34.9 32.9 24.1 42.3 48.5 50.0 22.6 24.4 22.5 24.4
# [211] 21.7 19.3 22.4 28.1 23.7 25.0 23.3 28.7 21.5 23.0 26.7 21.7 27.5 30.1
# [226] 50.0 37.6 31.6 46.7 31.5 24.3 31.7 41.7 48.3 29.0 24.0 25.1 31.5 23.7
# [241] 27.0 20.1 22.2 23.7 17.6 18.5 24.3 20.5 24.5 26.2 24.4 24.8 29.6 42.8
# [256] 20.9 44.0 50.0 36.0 30.1 33.8 43.1 48.8 31.0 36.5 22.8 30.7 50.0 43.5
# [271] 21.1 25.2 24.4 35.2 32.4 32.0 33.2 33.1 29.1 35.1 45.4 35.4 46.0 50.0
# [286] 22.0 20.1 23.2 22.3 24.8 28.5 37.3 27.9 23.9 21.7 28.6 27.1 20.3 22.5
# [301] 24.8 22.0 26.4 33.1 36.1 28.4 33.4 28.2 22.8 20.3 16.1 22.1 19.4 21.6
# [316] 16.2 17.8 19.8 23.1 21.0 23.8 23.1 20.4 18.5 25.0 24.6 23.0 22.2 19.3
# [331] 19.8 17.1 19.4 22.2 20.7 21.1 19.5 18.5 20.6 19.0 18.7 32.7 16.5 23.9
# [346] 17.5 17.2 23.1 24.5 26.6 22.9 24.1 18.6 30.1 18.2 20.6 17.8 21.7 22.7
# [361] 25.0 19.9 20.8 16.8 21.9 27.5 21.9 23.1 50.0 50.0 50.0 50.0 50.0 13.8
```

```
w[,1]#w数据的第一列
```

```
#   [1] 24.0 21.6 34.7 33.4 36.2 28.7 22.9 22.1 16.5 18.9 15.0 18.9 21.7 20.4
#  [16] 19.9 23.1 17.5 20.2 18.2 13.6 19.6 15.2 14.5 15.6 13.9 16.6 14.8 18.4
#  [31] 12.7 14.5 13.2 13.1 13.5 18.9 20.0 21.0 24.2 30.8 34.9 26.6 25.3 24.7
#  [46] 19.3 20.0 16.6 14.4 19.4 19.7 20.5 25.0 23.4 18.9 35.4 24.7 31.6 23.3
#  [61] 18.7 16.0 22.2 25.0 33.0 23.5 19.4 22.0 17.4 20.9 24.2 21.7 22.8 23.4
#  [76] 21.4 20.0 20.8 21.2 20.3 28.0 23.9 24.8 22.9 23.9 26.6 22.5 22.2 23.6
#  [91] 22.6 22.0 22.9 25.0 20.6 28.4 21.4 38.7 43.8 33.2 27.5 26.5 18.6 19.3
# [106] 19.5 19.5 20.4 19.8 19.4 21.7 22.8 18.8 18.7 18.5 18.3 21.2 19.2 20.4
# [121] 22.0 20.3 20.5 17.3 18.8 21.4 15.7 16.2 18.0 14.3 19.2 19.6 23.0 18.4
# [136] 18.1 17.4 17.1 13.3 17.8 14.0 14.4 13.4 15.6 11.8 13.8 15.6 14.6 17.8
# [151] 21.5 19.6 15.3 19.4 17.0 15.6 13.1 41.3 24.3 23.3 27.0 50.0 50.0 50.0
# [166] 25.0 50.0 23.8 23.8 22.3 17.4 19.1 23.1 23.6 22.6 29.4 23.2 24.6 29.9
# [181] 39.8 36.2 37.9 32.5 26.4 29.6 50.0 32.0 29.8 34.9 33.0 30.5 36.4 31.1
# [196] 50.0 33.3 30.3 34.6 34.9 32.9 24.1 42.3 48.5 50.0 22.6 24.4 22.5 24.4
# [211] 21.7 19.3 22.4 28.1 23.7 25.0 23.3 28.7 21.5 23.0 26.7 21.7 27.5 30.1
# [226] 50.0 37.6 31.6 46.7 31.5 24.3 31.7 41.7 48.3 29.0 24.0 25.1 31.5 23.7
# [241] 27.0 20.1 22.2 23.7 17.6 18.5 24.3 20.5 24.5 26.2 24.4 24.8 29.6 42.8
# [256] 20.9 44.0 50.0 36.0 30.1 33.8 43.1 48.8 31.0 36.5 22.8 30.7 50.0 43.5
# [271] 21.1 25.2 24.4 35.2 32.4 32.0 33.2 33.1 29.1 35.1 45.4 35.4 46.0 50.0
# [286] 22.0 20.1 23.2 22.3 24.8 28.5 37.3 27.9 23.9 21.7 28.6 27.1 20.3 22.5
# [301] 24.8 22.0 26.4 33.1 36.1 28.4 33.4 28.2 22.8 20.3 16.1 22.1 19.4 21.6
# [316] 16.2 17.8 19.8 23.1 21.0 23.8 23.1 20.4 18.5 25.0 24.6 23.0 22.2 19.3
# [331] 19.8 17.1 19.4 22.2 20.7 21.1 19.5 18.5 20.6 19.0 18.7 32.7 16.5 23.9
# [346] 17.5 17.2 23.1 24.5 26.6 22.9 24.1 18.6 30.1 18.2 20.6 17.8 21.7 22.7
# [361] 25.0 19.9 20.8 16.8 21.9 27.5 21.9 23.1 50.0 50.0 50.0 50.0 50.0 13.8
# [376] 15.0 13.9 13.3 13.1 10.2 10.4 10.9 11.3 12.3  8.8  7.2 10.5  7.4 10.2
# [391] 15.1 23.2  9.7 13.8 12.7 13.1 12.5  8.5  5.0  6.3  5.6  7.2 12.1  8.3
# [406]  5.0 11.9 27.9 17.2 27.5 15.0 17.2 17.9 16.3  7.0  7.2  7.5 10.4  8.8
# [421] 16.7 14.2 20.8 13.4 11.7  8.3 10.2 10.9 11.0  9.5 14.5 14.1 16.1 14.3
# [436] 13.4  9.6  8.2  8.4 12.8 10.5 17.1 14.8 15.4 10.8 11.8 14.9 12.6 14.1
# [451] 13.4 15.2 16.1 17.8 14.4 14.1 12.7 13.5 14.9 20.0 16.4 17.7 19.5 20.2
# [466] 19.9 19.0 19.1 19.1 20.1 19.9 19.6 23.2 29.8 13.8 13.3 16.7 12.0 14.6
```

Boston房价数据

```
w[,"cmedv"]#w数据的变量名为cmedv的列
```

```
#   [1] 24.0 21.6 34.7 33.4 36.2 28.7 22.9 22.1 16.5 18.9 15.0 18.9 21.7 20.4
#  [16] 19.9 23.1 17.5 20.2 18.2 13.6 19.6 15.2 14.5 15.6 13.9 16.6 14.8 18.4
#  [31] 12.7 14.5 13.2 13.1 13.5 18.9 20.0 21.0 24.2 30.8 34.9 26.6 25.3 24.7
#  [46] 19.3 20.0 16.6 14.4 19.4 19.7 20.5 25.0 23.4 18.9 35.4 24.7 31.6 23.3
#  [61] 18.7 16.0 22.2 25.0 33.0 23.5 19.4 22.0 17.4 20.9 24.2 21.7 22.8 23.4
#  [76] 21.4 20.0 20.8 21.2 20.3 28.0 23.9 24.8 22.9 23.9 26.6 22.5 22.2 23.6
#  [91] 22.6 22.0 22.9 25.0 20.6 28.4 21.4 38.7 43.8 33.2 27.5 26.5 18.6 19.3
# [106] 19.5 19.5 20.4 19.8 19.4 21.7 22.8 18.8 18.7 18.5 18.3 21.2 19.2 20.4
# [121] 22.0 20.3 20.5 17.3 18.8 21.4 15.7 16.2 18.0 14.3 19.2 19.6 23.0 18.4
# [136] 18.1 17.4 17.1 13.3 17.8 14.0 14.4 13.4 15.6 11.8 13.8 15.6 14.6 17.8
# [151] 21.5 19.6 15.3 19.4 17.0 15.6 13.1 41.3 24.3 23.3 27.0 50.0 50.0 50.0
# [166] 25.0 50.0 23.8 23.8 22.3 17.4 19.1 23.1 23.6 22.6 29.4 23.2 24.6 29.9
# [181] 39.8 36.2 37.9 32.5 26.4 29.6 50.0 32.0 29.8 34.9 33.0 30.5 36.4 31.1
# [196] 50.0 33.3 30.3 34.6 34.9 32.9 24.1 42.3 48.5 50.0 22.6 24.4 22.5 24.4
# [211] 21.7 19.3 22.4 28.1 23.7 25.0 23.3 28.7 21.5 23.0 26.7 21.7 27.5 30.1
# [226] 50.0 37.6 31.6 46.7 31.5 24.3 31.7 41.7 48.3 29.0 24.0 25.1 31.5 23.7
# [241] 27.0 20.1 22.2 23.7 17.6 18.5 24.3 20.5 24.5 26.2 24.4 24.8 29.6 42.8
# [256] 20.9 44.0 50.0 36.0 30.1 33.8 43.1 48.8 31.0 36.5 22.8 30.7 50.0 43.5
# [271] 21.1 25.2 24.4 35.2 32.4 32.0 33.2 33.1 29.1 35.1 45.4 35.4 46.0 50.0
# [286] 22.0 20.1 23.2 22.3 24.8 28.5 37.3 27.9 23.9 21.7 28.6 27.1 20.3 22.5
# [301] 24.8 22.0 26.4 33.1 36.1 28.4 33.4 28.2 22.8 20.3 16.1 22.1 19.4 21.6
# [316] 16.2 17.8 19.8 23.1 21.0 23.8 23.1 20.4 18.5 25.0 24.6 23.0 22.2 19.3
# [331] 19.8 17.1 19.4 22.2 20.7 21.1 19.5 18.5 20.6 19.0 18.7 32.7 16.5 23.9
# [346] 17.5 17.2 23.1 24.5 26.6 22.9 24.1 18.6 30.1 18.2 20.6 17.8 21.7 22.7
# [361] 25.0 19.9 20.8 16.8 21.9 27.5 21.9 23.1 50.0 50.0 50.0 50.0 50.0 13.8
```

```r
w[100:110,"cmedv"]
```

```
#  [1] 33.2 27.5 26.5 18.6 19.3 20.1 19.5 19.5 20.4 19.8 19.4
```

```r
class(w$cmedv)#变量的类型
```

```
# [1] "numeric"
```

```r
summary(w$cmedv)#数据的概况
```

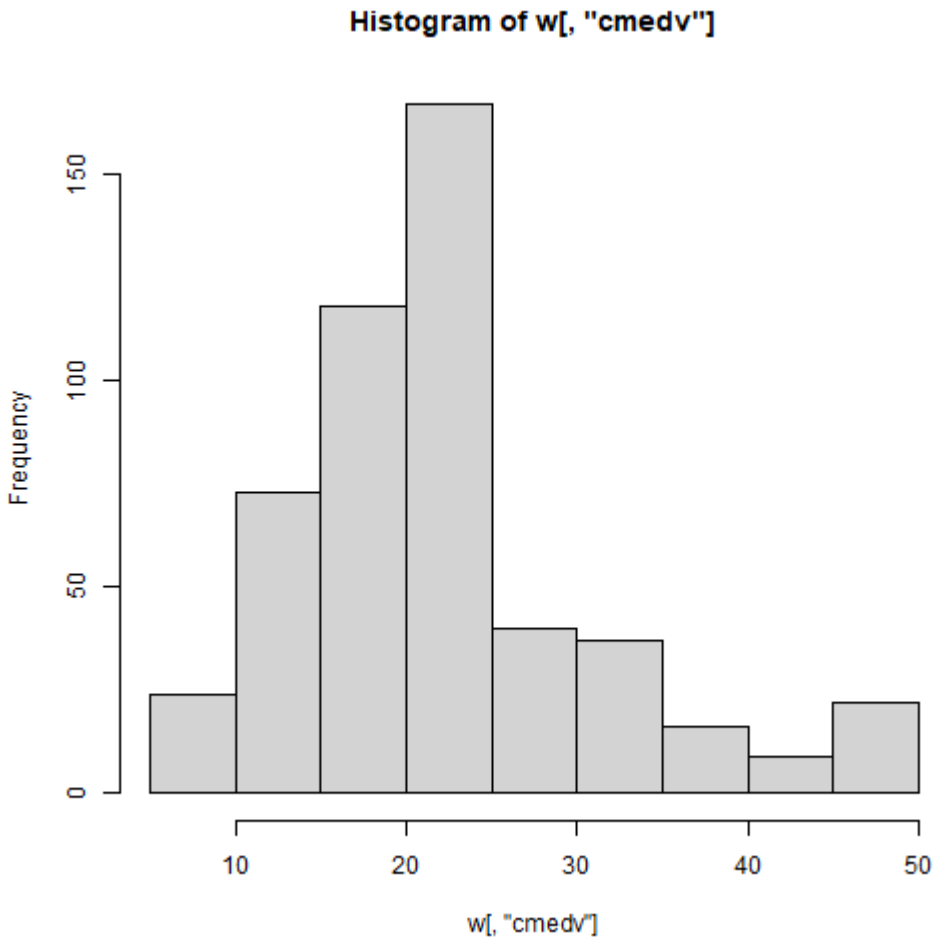
```
#    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#    5.00   17.02   21.20   22.53   25.00   50.00
```

```r
var(w$cmedv)
```

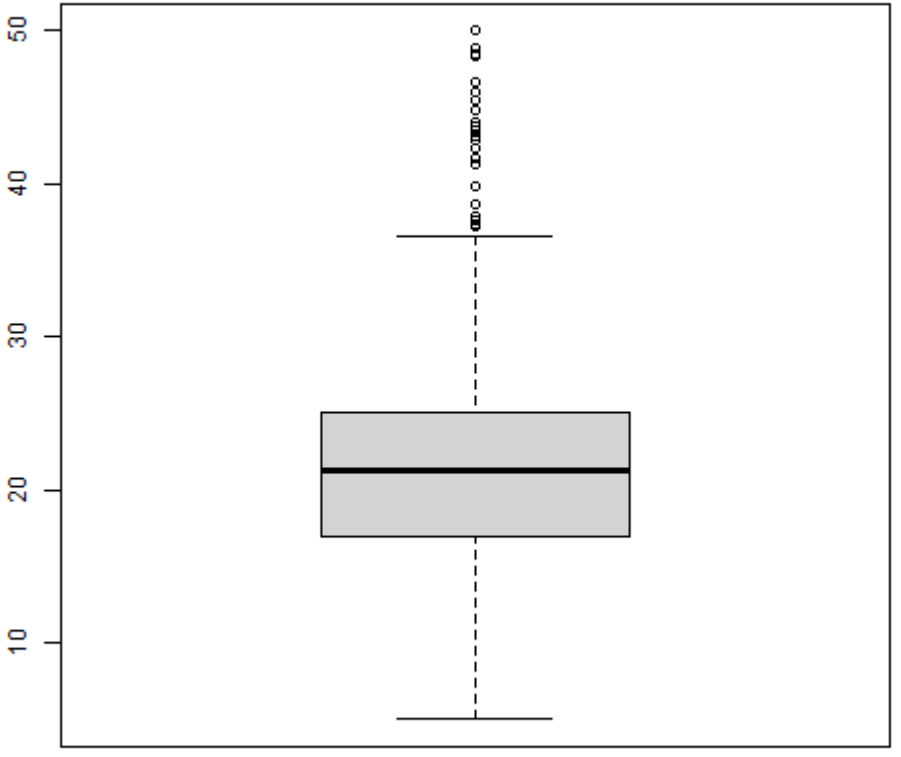
```
# [1] 84.31235
```

```r
# mean均值, quantile分位数, max最大值, min最小值, var方差, sd标准差
```

```
hist(w[,"cmedv"])#直方图
```



```r
boxplot(w[,"cmedv"])#箱线图
```




```r
#quantile(w$cmedv)
```

```
class(w$chas)#查看变量的类型
```

```
# [1] "integer"
```

```
summary(w$chas)#得到的不合理
```

```
#    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 0.00000 0.00000 0.00000 0.06917 0.00000 1.00000
```

```
table(w$chas)
```

```
#
#   0   1
# 471  35
```

```
w$chas<-factor(w$chas)#将其变为因子型变量
class(w$chas)
```

```
# [1] "factor"
```

```
summary(w$chas)
```

```
#   0   1
# 471  35
```

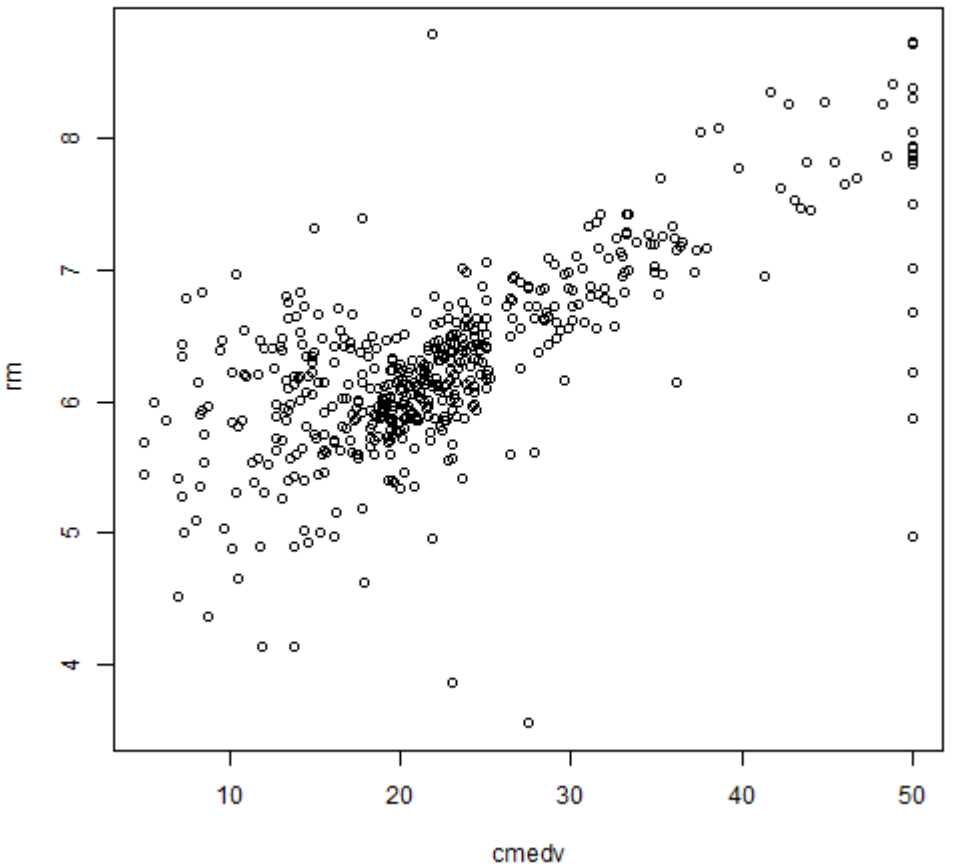
```
table(w$chas)
```

```
#
#   0   1
# 471  35
```

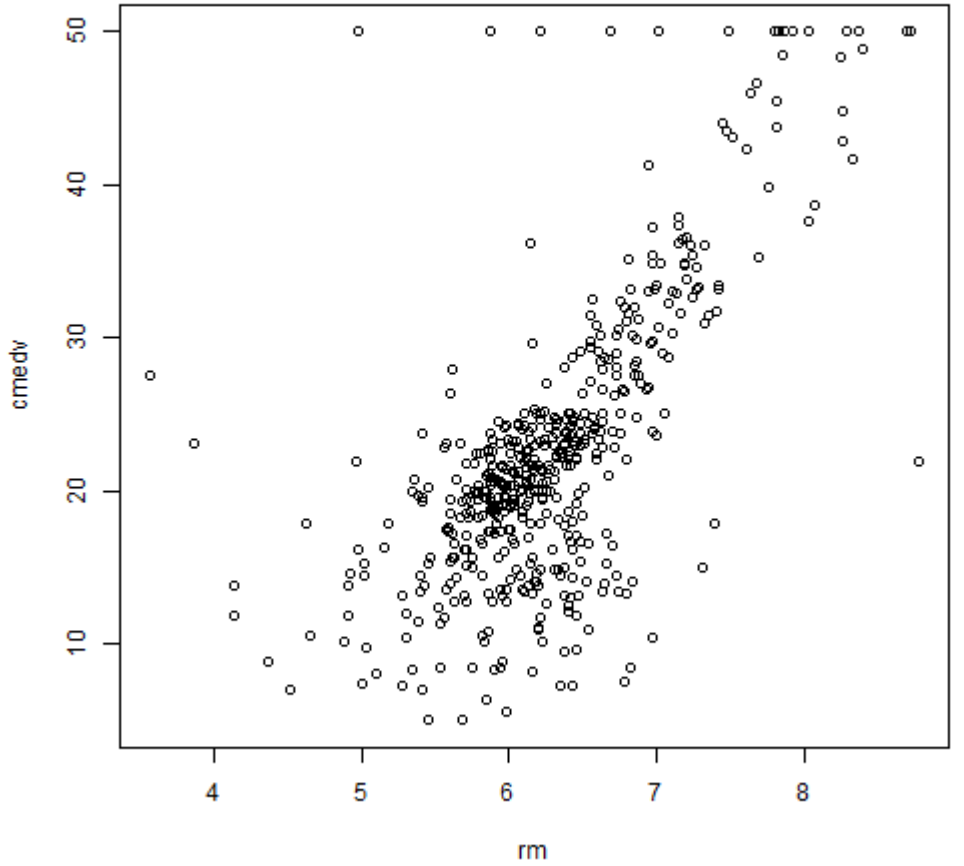
```r
summary(w[,c("cmedv", "rm")])
```

```
#       cmedv             rm
#  Min.   : 5.00   Min.   :3.561
#  1st Qu.:17.02   1st Qu.:5.886
#  Median :21.20   Median :6.208
#  Mean   :22.53   Mean   :6.285
#  3rd Qu.:25.00   3rd Qu.:6.623
#  Max.   :50.00   Max.   :8.780
```

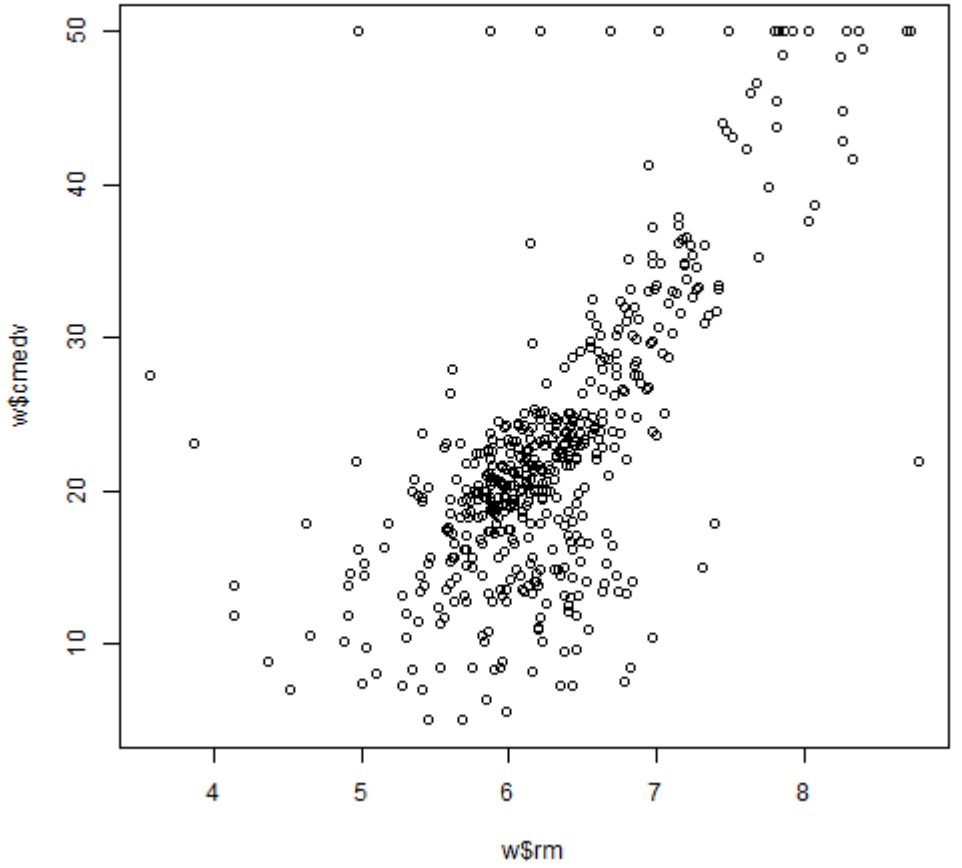
```
rm.cmedv.data<-w[,c("cmedv", "rm")]#cmedv和rm两列数据
plot(rm.cmedv.data);#cmedv和rm的散点图
```



```
plot(w[,c("rm", "cmedv")])#cmedv和rm的散点图
```



```
plot(w$rm, w$cmedv)#cmedv和rm的散点图
```



```r
# install.packages("GGally")#安装GGally软件包
# install.packages("corrplot")#安装GGally软件包
library(GGally)#加载GGally
```

```
# Warning: package 'GGally' was built under R version 4.0.4

# Loading required package: ggplot2

# Registered S3 method overwritten by 'GGally':
#   method from
#   +.gg   ggplot2
```

```r
library(corrplot)#加载corrplot
```

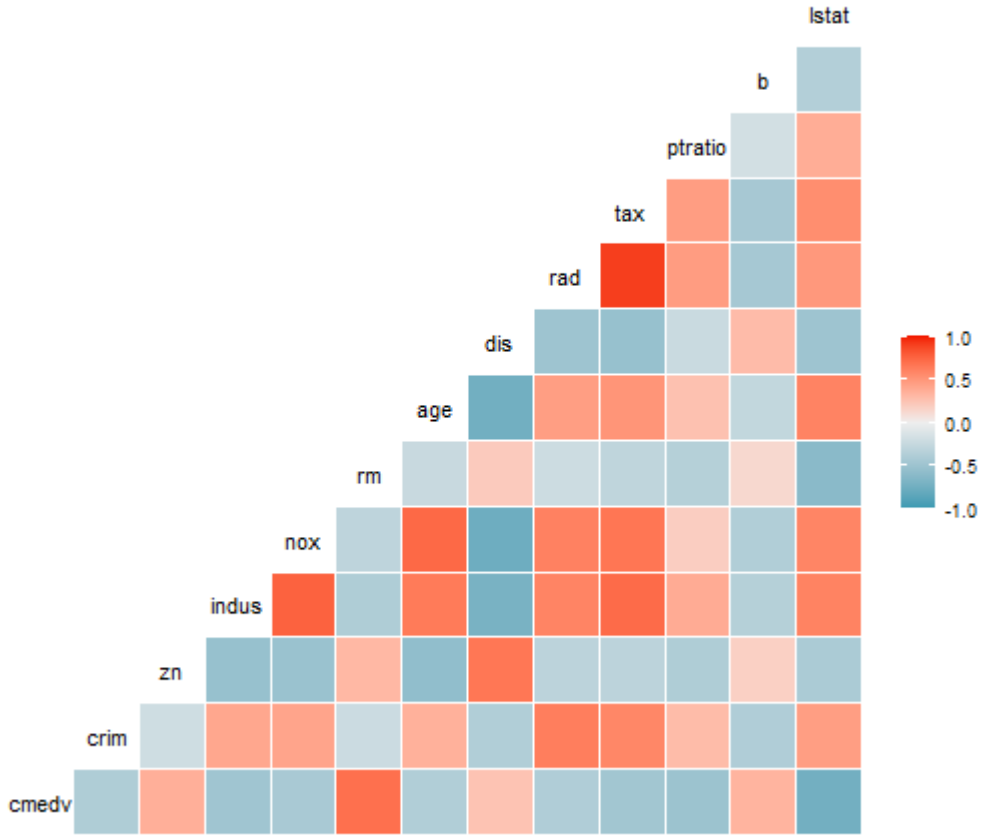
```
# Warning: package 'corrplot' was built under R version 4.0.4

# corrplot 0.84 loaded
```

```r
ggpairs(w)
```

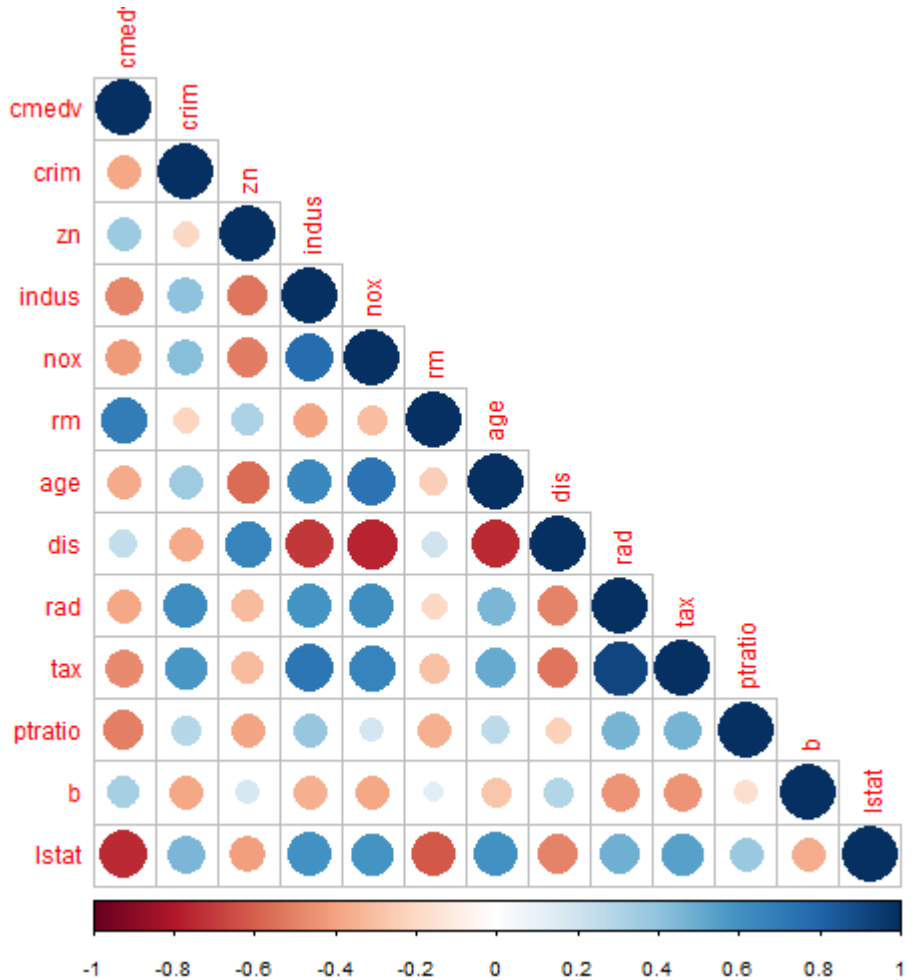
```
# `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
# `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
# `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
# `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
# `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
# `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
# `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
# `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
# `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
# `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
# `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
# `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
# `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
ggcorr(w)
```

```
# Warning in ggcorr(w): data in column(s) 'chas' are not numeric and were ign
```



```
#因为第5列chas是因子型变量，无法计算相关系数，因而将其去掉
corrplot(corr=cor(w[,-5]), type="lower")
```



```
summary(w)
```

```
#      cmedv            crim                zn              indus        chas
#  Min.   : 5.00   Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   0:471
#  1st Qu.:17.02   1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1: 35
#  Median :21.20   Median : 0.25651   Median :  0.00   Median : 9.69
#  Mean   :22.53   Mean   : 3.61352   Mean   : 11.36   Mean   :11.14
#  3rd Qu.:25.00   3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10
#  Max.   :50.00   Max.   :88.97620   Max.   :100.00   Max.   :27.74
#       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            b
#  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
#  Min.   : 1.73
#  1st Qu.: 6.95
#  Median :11.36
#  Mean   :12.65
#  3rd Qu.:16.95
#  Max.   :37.97
```

```
a=lm(cmedv~., data=w)#以cmedv为因变量，其他为自变量，拟合最小二乘回归
a$coef#最小二乘估计
```

```
#   (Intercept)          crim            zn         indus          chas1
#  3.637189e+01 -1.062004e-01  4.772368e-02  2.325237e-02  2.691727e+00
#           nox            rm           age           dis           rad
# -1.774262e+01  3.789395e+00  5.749168e-04 -1.501794e+00  3.037606e-01
#           tax       ptratio             b         lstat
# -1.270462e-02 -9.239118e-01  9.228445e-03 -5.306619e-01
```

```
coef(a)#最小二乘估计
```

```
#   (Intercept)          crim            zn         indus          chas1
#  3.637189e+01 -1.062004e-01  4.772368e-02  2.325237e-02  2.691727e+00
#           nox            rm           age           dis           rad
# -1.774262e+01  3.789395e+00  5.749168e-04 -1.501794e+00  3.037606e-01
#           tax       ptratio             b         lstat
# -1.270462e-02 -9.239118e-01  9.228445e-03 -5.306619e-01
```

```
summary(a)
```

```
#
# Call:
# lm(formula = cmedv ~ ., data = w)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -15.5651  -2.6908  -0.5352   1.8446  26.1319
#
# Coefficients:
#               Estimate Std. Error t value Pr(>|t|)
# (Intercept)  3.637e+01  5.058e+00   7.191 2.40e-12 ***
# crim        -1.062e-01  3.257e-02  -3.261 0.001189 **
# zn           4.772e-02  1.360e-02   3.508 0.000493 ***
# indus        2.325e-02  6.094e-02   0.382 0.702970
# chas1        2.692e+00  8.539e-01   3.152 0.001718 **
# nox         -1.774e+01  3.785e+00  -4.687 3.59e-06 ***
# rm           3.789e+00  4.142e-01   9.149  < 2e-16 ***
# age          5.749e-04  1.309e-02   0.044 0.964989
# dis         -1.502e+00  1.977e-01  -7.598 1.53e-13 ***
# rad          3.038e-01  6.575e-02   4.620 4.91e-06 ***
# tax         -1.270e-02  3.727e-03  -3.409 0.000706 ***
# ptratio     -9.239e-01  1.297e-01  -7.126 3.70e-12 ***
# b            9.228e-03  2.662e-03   3.467 0.000573 ***
# lstat       -5.307e-01  5.026e-02 -10.558  < 2e-16 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 4.703 on 492 degrees of freedom
# Multiple R-squared:  0.7444,    Adjusted R-squared:  0.7377
# F-statistic: 110.2 on 13 and 492 DF,  p-value: < 2.2e-16
```

蟹蟹

本幻灯片由 R 包 **xaringan** 生成；

查克拉来自于 remark.js、**knitr**、以及 R Markdown。