Homework 2

Author: Dong Bin

email: bindong2@illinois.edu

Question 1

Part 1

```
rm(list = ls())
library(mlbench)
data(BostonHousing2)
BH = BostonHousing2[, !(colnames(BostonHousing2) %in% c("medv", "town", "tract"))]

# Get some basic informations
dim(BH)
```

```
## [1] 506  16
```

```
names(BH)
```

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

```
# Fit a LM model
full.model <- lm(cmedv~., data = BH)
summary(full.model)
```

```
##
## Call:
## lm(formula = cmedv ~ ., data = BH)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.5831  -2.7643  -0.5994   1.7482  26.0822
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.350e+02  3.032e+02  -1.435 0.152029
## lon         -3.935e+00  3.372e+00  -1.167 0.243770
## lat          4.495e+00  3.669e+00   1.225 0.221055
## crim        -1.045e-01  3.261e-02  -3.206 0.001436 **
## zn           4.657e-02  1.374e-02   3.390 0.000755 ***
## indus        1.524e-02  6.175e-02   0.247 0.805106
## chas1        2.578e+00  8.650e-01   2.980 0.003024 **
## nox         -1.582e+01  4.005e+00  -3.951 8.93e-05 ***
## rm           3.754e+00  4.166e-01   9.011  < 2e-16 ***
```

```
## age          2.468e-03  1.335e-02   0.185 0.853440
## dis         -1.400e+00  2.088e-01  -6.704 5.61e-11 ***
## rad          3.067e-01  6.658e-02   4.607 5.23e-06 ***
## tax         -1.289e-02  3.727e-03  -3.458 0.000592 ***
## ptratio     -8.771e-01  1.363e-01  -6.436 2.92e-10 ***
## b            9.176e-03  2.663e-03   3.446 0.000618 ***
## lstat       -5.374e-01  5.042e-02 -10.660  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.7 on 490 degrees of freedom
## Multiple R-squared:  0.7458, Adjusted R-squared:  0.738
## F-statistic: 95.82 on 15 and 490 DF,  p-value: < 2.2e-16
```

The most significant variables according to P value are: 1. rm

2. lstat

Part 2

```
p <- dim(BH)[2]

test <- step(full.model, k = log(p))
```

```
## Start:  AIC=1594.28
## cmedv ~ lon + lat + crim + zn + indus + chas + nox + rm + age +
##     dis + rad + tax + ptratio + b + lstat
##
##           Df Sum of Sq   RSS    AIC
## - age      1      0.75 10826 1591.5
## - indus    1      1.35 10826 1591.6
## - lon      1     30.09 10855 1592.9
## - lat      1     33.17 10858 1593.0
## <none>                 10825 1594.3
## - chas     1    196.21 11021 1600.6
## - crim     1    227.01 11052 1602.0
## - zn       1    253.89 11079 1603.2
## - b        1    262.35 11087 1603.6
## - tax      1    264.16 11089 1603.7
## - nox      1    344.85 11170 1607.4
## - rad      1    468.79 11294 1613.0
## - ptratio  1    915.13 11740 1632.6
## - dis      1    992.75 11818 1635.9
## - rm       1   1793.75 12619 1669.1
## - lstat    1   2510.61 13336 1697.0
##
## Step:  AIC=1591.54
## cmedv ~ lon + lat + crim + zn + indus + chas + nox + rm + dis +
##     rad + tax + ptratio + b + lstat
##
##           Df Sum of Sq   RSS    AIC
```

```
## - indus    1      1.42 10827 1588.8
## - lon      1     29.36 10855 1590.1
## - lat      1     33.69 10859 1590.3
## <none>                 10826 1591.5
## - chas     1    199.53 11025 1598.0
## - crim     1    227.26 11053 1599.3
## - zn       1    253.44 11079 1600.5
## - tax      1    263.54 11089 1600.9
## - b        1    264.77 11090 1601.0
## - nox      1    352.01 11178 1605.0
## - rad      1    468.06 11294 1610.2
## - ptratio  1    914.57 11740 1629.8
## - dis      1   1122.29 11948 1638.7
## - rm       1   1905.55 12731 1670.8
## - lstat    1   2804.14 13630 1705.3
##
## Step:  AIC=1588.83
## cmedv ~ lon + lat + crim + zn + chas + nox + rm + dis + rad +
##     tax + ptratio + b + lstat
##
##           Df Sum of Sq   RSS    AIC
## - lon      1     32.13 10859 1587.6
## - lat      1     33.34 10860 1587.6
## <none>                 10827 1588.8
## - chas     1    203.29 11030 1595.5
## - crim     1    228.39 11056 1596.6
## - zn       1    252.83 11080 1597.7
## - b        1    263.94 11091 1598.2
## - tax      1    303.69 11131 1600.1
## - nox      1    372.59 11200 1603.2
## - rad      1    495.77 11323 1608.7
## - ptratio  1    929.67 11757 1627.7
## - dis      1   1173.09 12000 1638.1
## - rm       1   1915.38 12742 1668.5
## - lstat    1   2813.97 13641 1703.0
##
## Step:  AIC=1587.56
## cmedv ~ lat + crim + zn + chas + nox + rm + dis + rad + tax +
##     ptratio + b + lstat
##
##           Df Sum of Sq   RSS    AIC
## - lat      1     24.92 10884 1586.0
## <none>                 10859 1587.6
## - chas     1    235.62 11095 1595.7
## - crim     1    240.30 11100 1595.9
## - b        1    258.02 11117 1596.7
## - zn       1    281.72 11141 1597.8
## - tax      1    288.66 11148 1598.1
## - nox      1    504.45 11364 1607.8
## - rad      1    511.62 11371 1608.1
## - ptratio  1   1137.60 11997 1635.2
## - dis      1   1406.62 12266 1646.4
## - rm       1   1946.89 12806 1668.2
## - lstat    1   2810.02 13669 1701.2
```

```
##
## Step:  AIC=1585.95
## cmedv ~ crim + zn + chas + nox + rm + dis + rad + tax + ptratio +
##     b + lstat
##
##           Df Sum of Sq   RSS    AIC
## <none>                 10884 1586.0
## - chas     1    228.64 11113 1593.7
## - crim     1    237.49 11122 1594.1
## - b        1    265.68 11150 1595.4
## - zn       1    272.12 11156 1595.7
## - tax      1    287.68 11172 1596.4
## - rad      1    490.76 11375 1605.5
## - nox      1    538.23 11422 1607.6
## - ptratio  1   1132.44 12017 1633.3
## - dis      1   1502.93 12387 1648.6
## - rm       1   1940.06 12824 1666.2
## - lstat    1   2785.20 13669 1698.5
```

The variables removed from full model after stepwise regression with BIC criteria are: 1. age

2. indus
3. lon
4. lat

Part 3

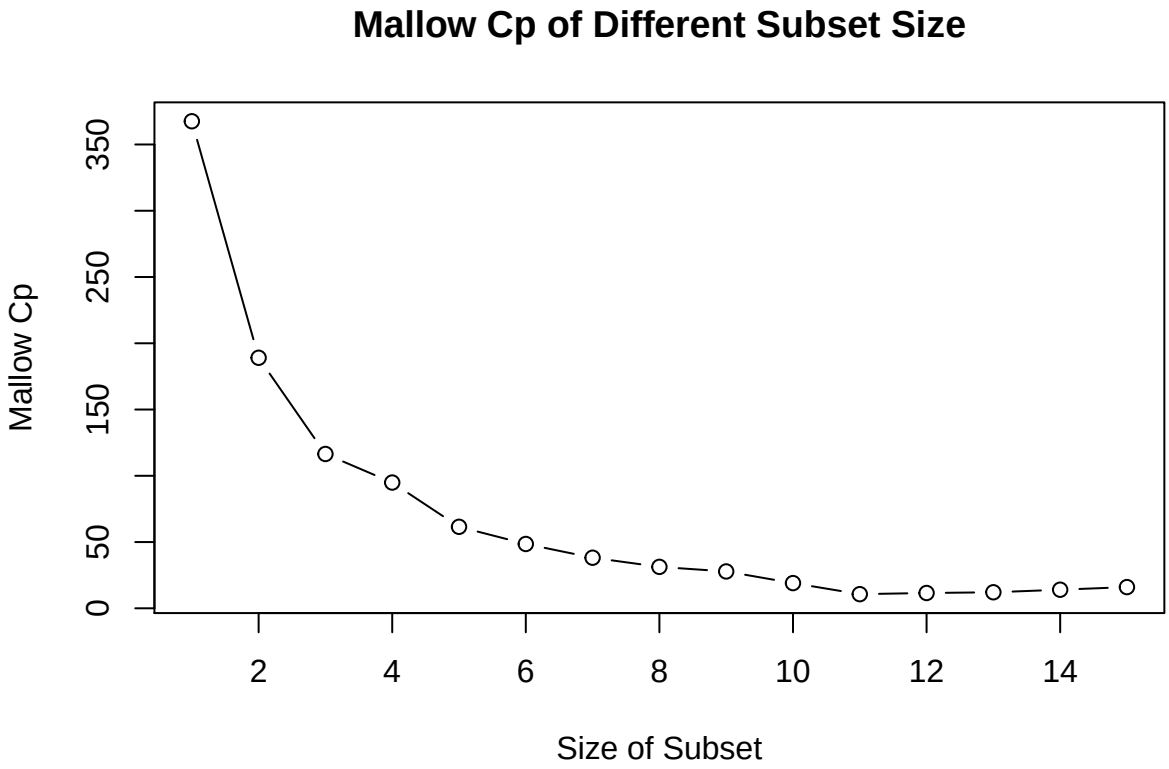
```
library(leaps)
b = regsubsets(cmedv~ ., data = BH, nvmax =p)
rs = summary(b)
rs$which
```

```
##    (Intercept)   lon   lat  crim    zn indus chas1   nox    rm   age   dis
## 1         TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 2         TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
## 3         TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
## 4         TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE
## 5         TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE
## 6         TRUE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE FALSE  TRUE
## 7         TRUE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE FALSE  TRUE
## 8         TRUE FALSE FALSE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE
## 9         TRUE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE FALSE  TRUE
## 10        TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE
## 11        TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE
## 12        TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE
## 13        TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE
## 14        TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
## 15        TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##      rad   tax ptratio     b lstat
```

```
## 1  FALSE FALSE   FALSE FALSE  TRUE
## 2  FALSE FALSE   FALSE FALSE  TRUE
## 3  FALSE FALSE    TRUE FALSE  TRUE
## 4  FALSE FALSE    TRUE FALSE  TRUE
## 5  FALSE FALSE    TRUE FALSE  TRUE
## 6  FALSE FALSE    TRUE FALSE  TRUE
## 7  FALSE FALSE    TRUE  TRUE  TRUE
## 8  FALSE FALSE    TRUE  TRUE  TRUE
## 9   TRUE  TRUE    TRUE  TRUE  TRUE
## 10  TRUE  TRUE    TRUE  TRUE  TRUE
## 11  TRUE  TRUE    TRUE  TRUE  TRUE
## 12  TRUE  TRUE    TRUE  TRUE  TRUE
## 13  TRUE  TRUE    TRUE  TRUE  TRUE
## 14  TRUE  TRUE    TRUE  TRUE  TRUE
## 15  TRUE  TRUE    TRUE  TRUE  TRUE
```

Part 4

```
row <- rs$which[1,]
names <- names(BH)
xlabel <- c(1:15)
plot(x = xlabel, y = rs$cp, type="b", main="Mallow Cp of Different Subset Size", xlab = "Size of Subset
```



```
rs$which[11,]
```

```
## (Intercept)         lon         lat        crim          zn       indus
##        TRUE       FALSE       FALSE        TRUE        TRUE       FALSE
##       chas1         nox          rm         age         dis         rad
##        TRUE        TRUE        TRUE       FALSE        TRUE        TRUE
##         tax     ptratio           b       lstat
##        TRUE        TRUE        TRUE        TRUE
```

The best model is when model size is 11. Remaining variables are: crim, zn, chas1, nox, rm, dis, rad, tax, ptratio, b, lstat

```
SubData <- BostonHousing2[, (colnames(BostonHousing2) %in% c("cmedv","crim", "zn", "chas1", "nox", "rm"
head(SubData)
```

```
##   cmedv    crim zn   nox    rm    dis rad tax ptratio      b lstat
## 1  24.0 0.00632 18 0.538 6.575 4.0900   1 296    15.3 396.90  4.98
## 2  21.6 0.02731  0 0.469 6.421 4.9671   2 242    17.8 396.90  9.14
## 3  34.7 0.02729  0 0.469 7.185 4.9671   2 242    17.8 392.83  4.03
## 4  33.4 0.03237  0 0.458 6.998 6.0622   3 222    18.7 394.63  2.94
## 5  36.2 0.06905  0 0.458 7.147 6.0622   3 222    18.7 396.90  5.33
## 6  28.7 0.02985  0 0.458 6.430 6.0622   3 222    18.7 394.12  5.21
```

```
summary(lm(cmedv~., data=SubData))
```

```
##
## Call:
## lm(formula = cmedv ~ ., data = SubData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.3325  -2.7562  -0.5958   1.9273  26.3651
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.524865   5.068796   7.206 2.17e-12 ***
## crim         -0.112317   0.032745  -3.430 0.000654 ***
## zn            0.046996   0.013528   3.474 0.000558 ***
## nox         -16.407119   3.525175  -4.654 4.18e-06 ***
## rm            3.821857   0.406256   9.408  < 2e-16 ***
## dis          -1.553761   0.185510  -8.376 5.70e-16 ***
## rad           0.312528   0.063230   4.943 1.06e-06 ***
## tax          -0.012976   0.003362  -3.860 0.000129 ***
## ptratio      -0.949052   0.128728  -7.373 7.09e-13 ***
## b             0.009643   0.002671   3.610 0.000338 ***
## lstat        -0.534008   0.047412 -11.263  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.738 on 495 degrees of freedom
## Multiple R-squared:  0.739,  Adjusted R-squared:  0.7337
## F-statistic: 140.2 on 10 and 495 DF,  p-value: < 2.2e-16
```

After removing insignificant variables, the most significant variables are :

1. rm
2. lstat

Question 2