2.

Homework1

2. a. $v \cdot w = 1\times2 - 1\times1 - 3\times1 = 2-1-3 = -2$

b. $-3 * w = -3\begin{bmatrix}2\\1\\-1\end{bmatrix} = \begin{bmatrix}-6\\-3\\1\end{bmatrix}$

c. $M * v = \begin{bmatrix}20 & 15 & 0\\5 & 25 & 10\\0 & 20 & 5\end{bmatrix}\begin{bmatrix}1\\-1\\3\end{bmatrix} = \begin{bmatrix}20-15+0\\5-25+30\\0-20+15\end{bmatrix} = \begin{bmatrix}5\\10\\-5\end{bmatrix}$

d. $M + N = \begin{bmatrix}20 & 15 & 0\\5 & 25 & 10\\0 & 20 & 5\end{bmatrix} + \begin{bmatrix}-20 & 5 & 10\\0 & -10 & 10\\5 & 20 & -5\end{bmatrix} = \begin{bmatrix}0 & 20 & 10\\5 & 15 & 20\\5 & 40 & 0\end{bmatrix}$

e. $M - N = \begin{bmatrix}20 & 15 & 0\\5 & 25 & 10\\0 & 20 & 5\end{bmatrix} - \begin{bmatrix}-20 & 5 & 10\\0 & -10 & 10\\5 & 20 & -5\end{bmatrix} = \begin{bmatrix}40 & 10 & -10\\5 & 35 & 0\\-5 & 0 & 10\end{bmatrix}$

f. $Z^T = \begin{bmatrix}1 & 1 & 1 & 1\\5 & -3 & 2 & 4\end{bmatrix}$

g. $Z^TZ = \begin{bmatrix}1 & 1 & 1 & 1\\5 & -3 & 2 & 4\end{bmatrix}\begin{bmatrix}1 & 5\\1 & -3\\1 & 2\\1 & 4\end{bmatrix} = \begin{bmatrix}1+1+1+1 & 5-3+2+4\\5-3+2+4 & 25+9+4+16\end{bmatrix} = \begin{bmatrix}4 & 8\\8 & 54\end{bmatrix}$

h. $(Z^TZ)^{-1} = \frac{1}{4\times54-64}\begin{bmatrix}54 & -8\\-8 & 4\end{bmatrix} = \frac{1}{152}\begin{bmatrix}54 & -8\\-8 & 4\end{bmatrix} = \begin{bmatrix}\frac{27}{76} & -\frac{1}{19}\\-\frac{1}{19} & \frac{1}{38}\end{bmatrix}$

i. $Z^TY = \begin{bmatrix}1 & 1 & 1 & 1\\5 & -3 & 2 & 4\end{bmatrix}\begin{bmatrix}2\\1\\-1\\3\end{bmatrix} = \begin{bmatrix}2+1-1+3\\10-3-2+12\end{bmatrix} = \begin{bmatrix}5\\17\end{bmatrix}$

j. $\beta = (Z^TZ)^{-1}Z^TY = \begin{bmatrix}\frac{27}{76} & -\frac{1}{19}\\-\frac{1}{19} & \frac{1}{38}\end{bmatrix}\begin{bmatrix}5\\17\end{bmatrix} = \begin{bmatrix}\frac{27\times5}{76} - \frac{17}{19}\\-\frac{5}{19} + \frac{17}{38}\end{bmatrix} = \begin{bmatrix}\frac{67}{76}\\\frac{7}{38}\end{bmatrix}$

k. $\det(Z^TZ) = 4\times54 - 64 = 216 - 64 = 152$

3.

Source Code:

```
#Problem 3
library(MASS)
#Create Matrix
Z <- matrix(c(1, 5, 1, -3, 1, 2, 1, 4),
        nrow = 4,
        ncol = 2,
        byrow = T)
Y <- matrix(c(2, 1, -1, 3),
        nrow = 4,
        ncol = 1,
        byrow = T)
M <- matrix(c(20, 15, 0, 5, 25, 10, 0, 20, 5),
        nrow = 3,
        ncol = 3,
        byrow = T)
```

```
N <- matrix(c(-20, 5, 10, 0, -10, 10, 5, 20, -5),
        nrow = 3,
        ncol = 3,
        byrow = T)
v <- matrix(c(1, -1, 3),
        nrow = 3,
        ncol = 1,
        byrow = T)
w <- matrix(c(2, 1, -1),
        nrow = 3,
        ncol = 1,
        byrow = T)
#a. v.w
crossprod(v, w)

#b. -3*w
-3 * w

#c. M * v
M %*% v

#d. M + N
M + N

#e. M - N
M - N

#f. Z(T)
Zt <- t(Z)
Zt

#g. Z(T)Z
ZtZ = Zt %*% Z
ZtZ

#h. (Z(T)Z)^(-1)
inverse <- fractions(ginv(ZtZ))
inverse

#i. Z(T)Y
ZtY <- Zt %*% Y
ZtY

#j. Beta
beta <- fractions(inverse %*% ZtY)
beta
```

#k. det(Z(T)Z)
detZtZ <- det(ZtZ)
detZtZ

#dataset x y
x <- c(5, -3, 2, 4)
y <- c(2, 1, -1, 3)

dataSet <- data.frame(x, y)
fit <- lm(y ~ x, data = dataSet)
summary(fit)

Output:

```
> #a. v.w
> crossprod(v, w)
     [,1]
[1,]   -2
>
> #b. -3*w
> -3 * w
     [,1]
[1,]   -6
[2,]   -3
[3,]    3
>
> #c. M * v
> M %*% v
     [,1]
[1,]    5
[2,]   10
[3,]   -5
> #f. Z(T)
> Zt <- t(Z)
> Zt
     [,1] [,2] [,3] [,4]
[1,]    1    1    1    1
[2,]    5   -3    2    4
>
> #g. Z(T)Z
> ZtZ = Zt %*% Z
> ZtZ
     [,1] [,2]
[1,]    4    8
[2,]    8   54
>
> #h. (Z(T)Z)^(-1)
> inverse <- fractions(ginv(ZtZ))
> inverse
      [,1]  [,2]
[1,] 27/76 -1/19
> #j. Beta
> beta <- fractions(inverse %*% ZtY)
> beta
      [,1]
[1,] 67/76
[2,]  7/38
>
> #k. det(Z(T)Z)
> detZtZ <- det(ZtZ)
> detZtZ
[1] 152
>
> #dataset x y
> x <- c(5, -3, 2, 4)
> y <- c(2, 1, -1, 3)
>
> dataSet <- data.frame(x, y)
> fit <- lm(y ~ x, data = dataSet)
> summary(fit)
```

```
> #d. M + N
> M + N
     [,1] [,2] [,3]
[1,]    0   20   10
[2,]    5   15   20
[3,]    5   40    0
>
> #e. M - N
> M - N
     [,1] [,2] [,3]
[1,]   40   10  -10
[2,]    5   35    0
[3,]   -5    0   10
>
> #f. Z(T)
> Zt <- t(Z)
> Zt
     [,1] [,2] [,3] [,4]
[1,]    1    1    1    1
> #h. (Z(T)Z)^(-1)
> inverse <- fractions(ginv(ZtZ))
> inverse
      [,1]  [,2]
[1,] 27/76 -1/19
[2,] -1/19  1/38
>
> #i. Z(T)Y
> ZtY <- Zt %*% Y
> ZtY
     [,1]
[1,]    5
[2,]   17
>
> #j. Beta
> beta <- fractions(inverse %*% ZtY)
> beta
      [,1]
[1,] 67/76
> summary(fit)

Call:
lm(formula = y ~ x, data = dataSet)

Residuals:
      1       2       3       4 
 0.1974  0.6711 -2.2500  1.3816 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.8816     1.1512   0.766    0.524
x             0.1842     0.3133   0.588    0.616

Residual standard error: 1.931 on 2 degrees of freedom
Multiple R-squared:  0.1474,	Adjusted R-squared:  -0.2789 
F-statistic: 0.3457 on 1 and 2 DF,  p-value: 0.6161
```

4.
a.
Source Code:
#Problem 4
#a.
head(mtcars)

A <- mtcars[c("cyl", "disp", "hp", "wt", "carb")]
Y <- mtcars[c("mpg")]

b.
Source Code:
#b.
A$count <- rep(1, nrow(A))
A <- A[c(6, 1, 2, 3, 4, 5)]

c.
Source Code:
#c.
A <- as.matrix(A)
Y <- as.matrix(Y)

d.
Source Code:
#d.
At <- t(A)
inverse <- fractions(ginv(At %*% A))
AtY <- At %*% Y
beta <- inverse %*% AtY
beta
Output:

```
> beta <- inverse %*% AtY
> beta
              mpg
[1,] 40.815359236
[2,] -1.291898563
[3,]  0.011485584
[4,] -0.020352893
[5,] -3.846949031
[6,] -0.006746893
```

e.
Source Code:
#e.
dataSet <- mtcars[c("cyl", "disp", "hp", "wt", "carb", "mpg")]
model <- lm(mpg ~ cyl + disp + hp + wt + carb, data = dataSet)
summary(model)
Output:

```
Call:
lm(formula = mpg ~ cyl + disp + hp + wt + carb, data = dataSet)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.0635 -1.4580 -0.4306  1.2927  5.8244 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 40.815359   3.025568  13.490    3e-13 ***
cyl         -1.291899   0.679227  -1.902  0.06830 .  
disp         0.011486   0.015375   0.747  0.46175    
hp          -0.020353   0.020062  -1.015  0.31968    
wt          -3.846949   1.192155  -3.227  0.00337 ** 
carb        -0.006747   0.574269  -0.012  0.99072    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.56 on 26 degrees of freedom
Multiple R-squared:  0.8486,	Adjusted R-squared:  0.8195 
F-statistic: 29.15 on 5 and 26 DF,  p-value: 7.056e-10
```

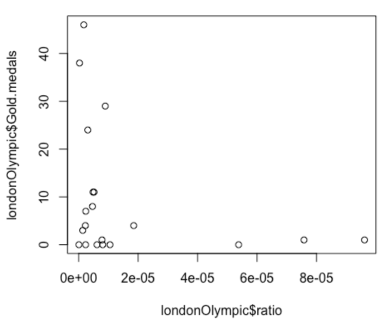
Comparing the output between d and e, I find that the matrix of beta is the same as the coefficients of the model between mpg and the other five attributes.

5.
Source Code:
londonOlympic <- read.csv2("/Users/Yiyang/Documents/CSC 424/Homework-1-DataFiles/olympics.csv", sep = ",", header = T)

londonOlympic$ratio <- (londonOlympic$Female.count + londonOlympic$Male.count)/londonOlympic$X2010.population

modelL <- lm(Gold.medals ~ ratio, data = londonOlympic)
summary(modelL)
plot(londonOlympic$ratio, londonOlympic$Gold.medals)

Output:



The point I am going to talk is between the athletics/population ratio and the amount of gold medals. As I see, almost all the country, the athletics ratio is similar, there are three outliers, which athletics ratio is larger than the others. Except these three countries, those countries which have the similar ratio, however, they don’t have the same amount of medals. The leader of the gold medals ranking is US, and second place is China, the third one is UK, I can find a thing, these three countries don’t have a high athletics ratio, the reason why they could get many medals, in my view, is the athletic practicing level, these countries spend a large part of money to support sports industry, so athlete could have a better environment to practice, and reach the best level to prepare matches.

6.
Source Code:
maple <- read.table("/Users/Yiyang/Documents/CSC 424/Homework-1-DataFiles/maple.txt", header = TRUE)
print(maple)
X1 <- maple$Latitude
X2 <- maple$JulyTemp
Y <- maple$LeafIndex

a.
Source Code:
#a.
m1 <- lm(Y ~ X1)
summary(m1)

Output:

```
> m1 <- lm(Y ~ X1)
> summary(m1)

Call:
lm(formula = Y ~ X1)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.2348 -0.8488  0.0773  1.0074  3.3305 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.66716    3.05202  -0.546    0.589    
X1           0.45369    0.07427   6.108 1.03e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.673 on 30 degrees of freedom
Multiple R-squared:  0.5543,	Adjusted R-squared:  0.5394 
F-statistic: 37.31 on 1 and 30 DF,  p-value: 1.031e-06
```

b.
Source Code:
#b.
m2 <- lm(Y ~ X2)
summary(m2)

Output:

```
> m2 <- lm(Y ~ X2)
> summary(m2)

Call:
lm(formula = Y ~ X2)

Residuals:
   Min     1Q Median     3Q    Max 
-3.254 -1.288  0.096  1.245  3.212 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 40.74297    4.45498   9.145 3.51e-10 ***
X2          -0.33318    0.06206  -5.368 8.23e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.789 on 30 degrees of freedom
Multiple R-squared:   0.49,	Adjusted R-squared:  0.473 
F-statistic: 28.82 on 1 and 30 DF,  p-value: 8.233e-06
```

c.

Source Code:
#c.
m3 <- lm(Y ~ X1 + X2)
summary(m3)

Output:

```
> m3 <- lm(Y ~ X1 + X2)
> summary(m3)

Call:
lm(formula = Y ~ X1 + X2)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.2082 -1.2363  0.1613  1.0551  3.2445 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 13.73184   11.42026   1.202   0.2389  
X1           0.31393    0.12388   2.534   0.0169 *
X2          -0.13524    0.09676  -1.398   0.1728  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.647 on 29 degrees of freedom
Multiple R-squared:  0.5824,	Adjusted R-squared:  0.5536 
F-statistic: 20.22 on 2 and 29 DF,  p-value: 3.167e-06
```

The coefficients of X1 in a) is 0.45369, in c) is 0.31393, X2 in b) is -0.33318, in c) is -0.13524. The coefficient of X1 is decreasing, X2 is increasing. Since X1 is the latitude, and X2 is the temperature, these two factors also influence each other, if latitude increase the temperature may decrease. This could be the reason why the coefficient changes.

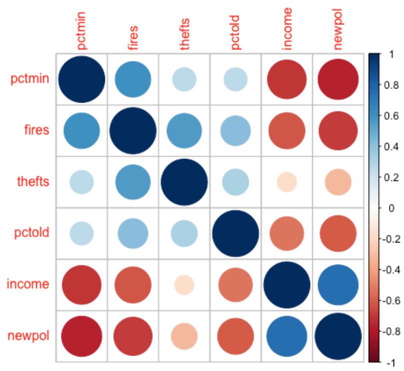
d.
I think p-value could detect this issue, the p-value of these two variables are 0.0169 for X1 and 0.1728 for X2.

7.
Source Code:
chicinsur <- read.table("/Users/Yiyang/Documents/CSC 424/Homework-1-DataFiles/chicinsur.txt", header = TRUE)
print(chicinsur)

a.
Source Code:
#a.
c <- data.frame(pctmin = chicinsur$pctmin,
            fires = chicinsur$fires,
            thefts = chicinsur$thefts,
            pctold = chicinsur$pctold,
            income = chicinsur$income,
            newpol = chicinsur$newpol)
C <- cor(c)
C
library(corrplot)
corrplot(C, method = "circle")

Output:

```
> C
          pctmin      fires     thefts     pctold     income     newpol
pctmin 1.0000000  0.5927956  0.2550647  0.2505118 -0.7037328 -0.7594196
fires  0.5927956  1.0000000  0.5562105  0.4122225 -0.6104481 -0.6864766
thefts 0.2550647  0.5562105  1.0000000  0.3176308 -0.1729226 -0.3116183
pctold 0.2505118  0.4122225  0.3176308  1.0000000 -0.5286695 -0.6057428
income -0.7037328 -0.6104481 -0.1729226 -0.5286695  1.0000000  0.7509780
newpol -0.7594196 -0.6864766 -0.3116183 -0.6057428  0.7509780  1.0000000
```



The coefficients matrix could support the prediction about the signs, and the correlation plot is very clear to show how the variables influence each other.

b.

Source Code:

c6 <- lm(newpol ~ fires + pctmin + thefts + pctold + income, data = chicinsur)

summary(c6)

Output:

```
Call:
lm(formula = newpol ~ fires + pctmin + thefts + pctold + income,
    data = chicinsur)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5235 -1.2134 -0.1544  1.0181  3.8096

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.0610686  2.8187818   4.279 0.000110 ***
fires       -0.1018544  0.0480110  -2.121 0.039972 *
pctmin      -0.0594738  0.0131806  -4.512  5.3e-05 ***
thefts       0.0135616  0.0162371   0.835 0.408436
pctold      -0.0643711  0.0158312  -4.066 0.000211 ***
income       0.0001164  0.0001804   0.645 0.522525
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.907 on 41 degrees of freedom
Multiple R-squared:  0.7939,    Adjusted R-squared:  0.7688
F-statistic: 31.59 on 5 and 41 DF,  p-value: 4.773e-13
```

i.

From the summary of the model, the adjusted R^2 is 0.7688, pctmin and pctold have the most significant correlation with newpol, fires has less significant correlation with newpol, the other variables don't have significant correlation with newpol.

ii.

From the report, thefts and income have coefficients that are significantly different form zero.

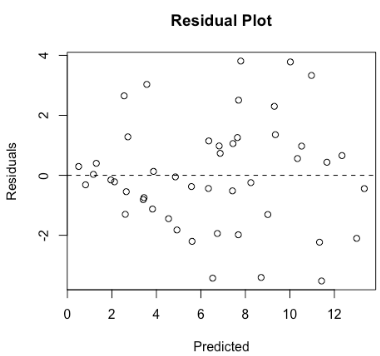
iii.
Almost all the predictors are different than suggested by their simple correlations. I think simple correlations are the relationship between two variables, but if put them together in to a multiple regression model, they will influence each other and then they change.

iv.
Source Code:

```
r = residuals(c6)
p = predict(c6)

plot(p, r, main = "Residual Plot",
        xlab = "Predicted",
        ylab = "Residuals")
abline(h = 0, lty = 2)
```

Output:



The problem I notice is that the predicted value and residual value are approaching to a straight line from 0 to 6, which means this plot is nearly heteroscedastic.

8.
Source Code:

```
housing <- read.table("/Users/Yiyang/Documents/CSC 424/housing.data", header = FALSE)
colnames(housing) <- c("CRIM", "ZN", "INDUS", "CHAS", "NOX", "RM", "AGE", "DIS", "RAD", "TAX",
"PTRATIO", "B", "LSTAT", "MEDV")
print(housing)

modelH <- lm(CRIM ~ ZN + INDUS + CHAS + NOX + RM + AGE + DIS + RAD + TAX + PTRATIO + B + LSTAT +
MEDV, data = housing)
summary(modelH)
```

Output:

```
> summary(modelH)

Call:
lm(formula = CRIM ~ ZN + INDUS + CHAS + NOX + RM + AGE + DIS +
    RAD + TAX + PTRATIO + B + LSTAT + MEDV, data = housing)

Residuals:
   Min     1Q Median     3Q    Max
-9.924 -2.120 -0.353  1.019 75.051

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.033228   7.234903   2.354 0.018949 *
ZN            0.044855   0.018734   2.394 0.017025 *
INDUS        -0.063855   0.083407  -0.766 0.444294
CHAS         -0.749134   1.180147  -0.635 0.525867
NOX         -10.313535   5.275536  -1.955 0.051152 .
RM            0.430131   0.612830   0.702 0.483089
AGE           0.001452   0.017925   0.081 0.935488
DIS          -0.987176   0.281817  -3.503 0.000502 ***
RAD           0.588209   0.088049   6.680 6.46e-11 ***
TAX          -0.003780   0.005156  -0.733 0.463793
PTRATIO      -0.271081   0.186450  -1.454 0.146611
B            -0.007538   0.003673  -2.052 0.040702 *
LSTAT         0.126211   0.075725   1.667 0.096208 .
MEDV         -0.198887   0.060516  -3.287 0.001087 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.439 on 492 degrees of freedom
Multiple R-squared:  0.454,	Adjusted R-squared:  0.4396
F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16
```

From the report, I get the R^2 is 0.454, and adjusted R^2 is 0.4396, the value is not that large, so I think this is not a very good model for these values. The utility for this model is going to predict the crime rate from different factors that may influence the crime rate. The estimated coefficients and Std. Errors of each variable are showing in the report, according to the report, DIS, RAD, ZN, B, MEDV, NOX and LSTAT have significant correlation, DIS and RAD have the highest significant correlation with CRIM.

b.

Source Code:

fwd.model <- step(lm(CRIM ~ 1, data = housing), direction = 'forward', scope = ~ ZN + INDUS + CHAS + NOX + RM + AGE + DIS + RAD + TAX + PTRATIO + B + LSTAT + MEDV)

Output:

```
Start:  AIC=2178.76
CRIM ~ 1

          Df Sum of Sq   RSS    AIC
+ RAD      1   14618.6 22745 1929.6
+ TAX      1   12689.1 24674 1970.8
+ LSTAT    1    7756.3 29607 2063.0
+ NOX      1    6621.4 30742 2082.1
+ INDUS    1    6176.5 31187 2089.3
+ MEDV     1    5633.6 31730 2098.1
+ B        1    5540.0 31823 2099.6
+ DIS      1    5385.9 31977 2102.0
+ AGE      1    4648.8 32714 2113.5
+ PTRATIO  1    3141.1 34222 2136.3
+ RM       1    1796.0 35567 2155.8
+ ZN       1    1501.5 35862 2160.0
<none>                 37363 2178.8
+ CHAS     1     116.7 37247 2179.2

Step:  AIC=1929.61
CRIM ~ RAD

          Df Sum of Sq   RSS    AIC
+ LSTAT    1   1103.70 21641 1906.4
+ MEDV     1    978.68 21766 1909.3
+ B        1    533.86 22211 1919.6
+ RM       1    302.58 22442 1924.8
+ DIS      1    244.47 22500 1926.1
+ AGE      1    214.87 22530 1926.8
+ CHAS     1     98.27 22646 1929.4
<none>                 22745 1929.6
+ NOX      1     88.51 22656 1929.6
+ INDUS    1     68.17 22676 1930.1
+ TAX      1     39.19 22705 1930.7
+ ZN       1      1.18 22743 1931.6
+ PTRATIO  1      0.03 22745 1931.6

Step:  AIC=1906.44
CRIM ~ RAD + LSTAT
```

```
Step:  AIC=1906.44
CRIM ~ RAD + LSTAT

          Df Sum of Sq   RSS    AIC
+ B        1   292.024 21349 1901.6
+ MEDV     1   138.075 21503 1905.2
+ ZN       1    97.116 21544 1906.2
<none>                 21641 1906.4
+ CHAS     1    64.158 21577 1906.9
+ INDUS    1    53.955 21587 1907.2
+ PTRATIO  1    43.610 21597 1907.4
+ NOX      1    26.311 21615 1907.8
+ DIS      1    22.336 21619 1907.9
+ RM       1     9.883 21631 1908.2
+ TAX      1     8.847 21632 1908.2
+ AGE      1     3.274 21638 1908.4

Step:  AIC=1901.56
CRIM ~ RAD + LSTAT + B

          Df Sum of Sq   RSS    AIC
+ MEDV     1   103.678 21245 1901.1
+ ZN       1    91.255 21258 1901.4
<none>                 21349 1901.6
+ INDUS    1    67.246 21282 1902.0
+ CHAS     1    53.606 21295 1902.3
+ NOX      1    42.746 21306 1902.5
+ PTRATIO  1    28.353 21320 1902.9
+ DIS      1    17.734 21331 1903.1
+ TAX      1    14.840 21334 1903.2
+ AGE      1     2.976 21346 1903.5
+ RM       1     2.346 21346 1903.5

Step:  AIC=1901.1
CRIM ~ RAD + LSTAT + B + MEDV

          Df Sum of Sq   RSS    AIC
+ ZN       1   110.203 21135 1900.5
+ PTRATIO  1   102.546 21143 1900.7
```

```
Step:  AIC=1901.1
CRIM ~ RAD + LSTAT + B + MEDV

          Df Sum of Sq   RSS    AIC
+ ZN       1   110.203 21135 1900.5
+ PTRATIO  1   102.546 21143 1900.7
<none>                 21245 1901.1
+ INDUS    1    77.439 21168 1901.2
+ RM       1    52.907 21192 1901.8
+ DIS      1    45.010 21200 1902.0
+ NOX      1    37.135 21208 1902.2
+ TAX      1    33.388 21212 1902.3
+ CHAS     1    29.184 21216 1902.4
+ AGE      1     0.109 21245 1903.1

Step:  AIC=1900.47
CRIM ~ RAD + LSTAT + B + MEDV + ZN

          Df Sum of Sq   RSS    AIC
+ DIS      1   276.074 20859 1895.8
<none>                 21135 1900.5
+ PTRATIO  1    62.129 21073 1901.0
+ RM       1    43.500 21092 1901.4
+ TAX      1    36.493 21098 1901.6
+ INDUS    1    27.998 21107 1901.8
+ AGE      1    23.189 21112 1901.9
+ CHAS     1    20.192 21115 1902.0
+ NOX      1     6.275 21129 1902.3

Step:  AIC=1895.81
CRIM ~ RAD + LSTAT + B + MEDV + ZN + DIS

          Df Sum of Sq   RSS    AIC
+ NOX      1   207.325 20652 1892.8
+ INDUS    1   199.695 20659 1893.0
+ TAX      1   123.327 20736 1894.8
<none>                 20859 1895.8
+ RM       1    33.598 20825 1897.0
+ CHAS     1    31.742 20827 1897.0
```

```
Step:  AIC=1895.81
CRIM ~ RAD + LSTAT + B + MEDV + ZN + DIS

          Df Sum of Sq   RSS    AIC
+ NOX      1   207.325 20652 1892.8
+ INDUS    1   199.695 20659 1893.0
+ TAX      1   123.327 20736 1894.8
<none>                 20859 1895.8
+ RM       1    33.598 20825 1897.0
+ CHAS     1    31.742 20827 1897.0
+ PTRATIO  1    29.805 20829 1897.1
+ AGE      1    13.250 20846 1897.5

Step:  AIC=1892.76
CRIM ~ RAD + LSTAT + B + MEDV + ZN + DIS + NOX

          Df Sum of Sq   RSS    AIC
+ PTRATIO  1   119.043 20533 1891.8
+ INDUS    1   104.802 20547 1892.2
<none>                 20652 1892.8
+ TAX      1    76.231 20575 1892.9
+ RM       1    36.002 20616 1893.9
+ CHAS     1    15.467 20636 1894.4
+ AGE      1     0.124 20652 1894.8

Step:  AIC=1891.83
CRIM ~ RAD + LSTAT + B + MEDV + ZN + DIS + NOX + PTRATIO

        Df Sum of Sq   RSS    AIC
<none>               20533 1891.8
+ INDUS  1    73.986 20459 1892.0
+ TAX    1    58.823 20474 1892.4
+ RM     1    32.583 20500 1893.0
+ CHAS   1    18.329 20514 1893.4
+ AGE    1     1.285 20531 1893.8
>
```

From the report, the model with smallest AIC model is CRIM ~ RAD + LSTAT + B + MEDV + ZN + DIS + NOX + PTRATIO, which means these factors may be the suitable factors for the final model.

c.
Source Code:
bwd.model <- step(lm(CRIM ~ ZN + INDUS + CHAS + NOX + RM + AGE
+ DIS + RAD + TAX + PTRATIO + B + LSTAT + MEDV, data = housing), direction = 'backward', scope = ~ 1)

Output:

```
Step:  AIC=1891.83
CRIM ~ ZN + NOX + DIS + RAD + PTRATIO + B + LSTAT + MEDV

          Df Sum of Sq   RSS    AIC
<none>                 20533 1891.8
- LSTAT    1     104.7 20637 1892.4
- PTRATIO  1     119.0 20652 1892.8
- B        1     198.4 20731 1894.7
- ZN       1     239.6 20772 1895.7
- NOX      1     296.6 20829 1897.1
- MEDV     1     430.2 20963 1900.3
- DIS      1     507.8 21040 1902.2
- RAD      1    4739.5 25272 1994.9
>
```

The final model I get from backward selection method is CRIM ~ ZN + NOX + DIS + RAD + PTRATIO + B + LSTAT + MEDV, which is actually same as the forward selection method. So it means this model is the best model for this problem.