HW1_solution

```
library(MASS)
library(glmnet)

data(Boston)
y <- Boston[, 1]
x <- Boston[, -1]
L.FUN <- function(y, p, si) log(dnorm(y, mean=p, sd=si))
n <- length(y)

grid <- expand.grid(rep(list(c(0, 1)), 13))[-1,]
colnames(grid) <- colnames(x)
rownames(grid) <- 1:8191

lambda <- 10^seq(0.8, -3, length=1000)
```

Question 1

```
L <- aic <- bic <- NULL

for (i in 1:nrow(grid)) {
  x.sub <- x[, grid[i,] == 1, drop=F]
  fit.lm <- lm(y ~., data = x.sub)
  p <- predict(fit.lm, x.sub)
  d <- sum(fit.lm$coef != 0)
  si <- sd(y)
  L[i] <- sum(L.FUN(y, p, si))
  aic[i] <- -2 * L[i] + d * 2
  bic[i] <- -2 * L[i] + d * log(n)
}

# the best AIC and BIC
AIC <- aic[which.min(aic)]
aic.pos <- grid[which.min(aic), ]
BIC <- bic[which.min(bic)]
bic.pos <- grid[which.min(bic), ]

res.Q1 <- data.frame(rbind(aic.pos, bic.pos), values=rbind(AIC, BIC))
row.names(res.Q1) <- c("AIC", "BIC")
res.Q1
```

```
##     zn indus chas nox rm age dis rad tax ptratio black lstat medv   values
## AIC  1     0    0   1  0   0   1   1   0       0     1     0    1 3402.502
## BIC  0     0    0   0  0   0   0   1   0       0     0     1    0 3418.908
```

Question 2

```
L <- aic <- bic <- NULL

for (i in 1:nrow(grid)) {
  x.sub <- x[, grid[i,] == 1, drop=F]
  fit.lm <- lm(y ~., data = x.sub)
  p <- predict(fit.lm, x.sub)
  d <- sum(fit.lm$coef != 0)
  si <- summary(fit.lm)$sigma
  L[i] <- sum(L.FUN(y, p, si))
  aic[i] <- -2*L[i] + d * 2
  bic[i] <- -2*L[i] + d * log(n)
}

# the best AIC and BIC
AIC <- aic[which.min(aic)]
aic.pos <- grid[which.min(aic), ]
BIC <- bic[which.min(bic)]
bic.pos <- grid[which.min(bic), ]

res.Q2 <- data.frame(rbind(aic.pos, bic.pos), values=rbind(AIC, BIC))
row.names(res.Q2) <- c("AIC", "BIC")
res.Q2
```

```
##     zn indus chas nox rm age dis rad tax ptratio black lstat medv   values
## AIC  1     0    0   1  0   0   1   1   0       1     1     1    1 3327.881
## BIC  0     0    0   0  0   0   0   1   0       0     1     1    0 3354.450
```

Question 3

```
L <- aic <- bic <- NULL

fit.ls <- glmnet(x, y, alpha=1, lambda = lambda)
p <- predict(fit.ls, as.matrix(x))

for (i in 1:length(lambda)) {
  d <- sum(coef(fit.ls)[,i] != 0)
  si <- sd(y)
  L[i] <- sum(L.FUN(y, p[,i], si))
  aic[i] <- -2*L[i] + d * 2
  bic[i] <- -2*L[i] + d * log(n)
}

# the best AIC and BIC
aic.pos <- which.min(aic)
bic.pos <- which.min(bic)
AIC <- ifelse(fit.ls$beta[, aic.pos] != 0, 1, 0)
BIC <- ifelse(fit.ls$beta[, bic.pos] != 0, 1, 0)

res.Q3 <- data.frame(rbind(AIC, BIC), values=rbind(min(aic), min(bic)))
res.Q3
```

```
##     zn indus chas nox rm age dis rad tax ptratio black lstat medv   values
## AIC  1     0    1   0  0   0   1   1   0       0     1     1    1 3407.440
## BIC  0     0    0   0  0   0   0   1   0       0     1     1    1 3430.591
```

Question 4

```
L <- aic <- bic <- NULL

fit.ls <- glmnet(x, y, alpha=1, lambda = lambda)
p <- predict(fit.ls, as.matrix(x))

for(i in 1:length(lambda)){
  fit.lm <- lm(y ~ ., data = data.frame(y, x[, (fit.ls$beta[, i] != 0)]))
  d <- sum(fit.lm$coef != 0)
  si <- summary(fit.lm)$sigma
  L[i] <- sum(L.FUN(y, p[,i], si))
  aic[i] <- -2*L[i] + d * 2
  bic[i] <- -2*L[i] + d * log(n)
}

# the best AIC and BIC
aic.pos <- which.min(aic)
bic.pos <- which.min(bic)
AIC <- ifelse(fit.ls$beta[, aic.pos] != 0, 1, 0)
BIC <- ifelse(fit.ls$beta[, bic.pos] != 0, 1, 0)

res.Q4 <- data.frame(rbind(AIC, BIC),values=rbind(min(aic), min(bic)))
res.Q4
```

```
##     zn indus chas nox rm age dis rad tax ptratio black lstat medv   values
## AIC  1     1    1   1  1   0   1   1   0       1     1     1    1 3332.155
## BIC  0     0    0   0  0   0   1   1   0       0     1     1    1 3365.955
```

Question 5

```
L <- aic <- bic <- NULL

fit.ls <- glmnet(x, y, alpha=1, lambda = lambda)
p <- predict(fit.ls, as.matrix(x))

for(i in 1:length(lambda)){
  d <- sum(coef(fit.ls)[,i] != 0)
  si <- sqrt((1 / (n - d)) * sum((y - p[,i])^2) )
  L[i] <- sum(L.FUN(y, p[,i], si))
  aic[i] <- -2*L[i] + d * 2
  bic[i] <- -2*L[i] + d * log(n)
}

# the best AIC and BIC
aic.pos <- which.min(aic)
bic.pos <- which.min(bic)
AIC <- ifelse(fit.ls$beta[, aic.pos] != 0, 1, 0)
BIC <- ifelse(fit.ls$beta[, bic.pos] != 0, 1, 0)

res.Q5 <- data.frame(rbind(AIC, BIC),values=rbind(min(aic), min(bic)))
res.Q5
```

```
##     zn indus chas nox rm age dis rad tax ptratio black lstat medv   values
## AIC  1     1    1   1  1   0   1   1   0       1     1     1    1 3332.176
## BIC  0     0    0   0  0   0   1   1   0       0     1     1    1 3365.980
```

Question 6

```
set.seed(4321)
tran <- sample(nrow(x), 400)
test <- setdiff(1:nrow(x), tran)

tr.x <- x[tran, ]; tr.y <- y[tran]
te.x <- x[test, ]; te.y <- y[test]
n <- length(tr.y)
```

M1 - M4

```
L <- aic <- bic <- TE1 <- matrix(0, nrow(grid), 2)

for (i in 1:nrow(grid)) {
  trx.sub <- tr.x[, grid[i,] == 1, drop=F]
  fit.lm <- lm(tr.y ~., data = trx.sub)
  tr.p <- predict(fit.lm, tr.x)
  te.p <- predict(fit.lm, te.x)
  d <- sum(fit.lm$coef != 0)

  s1 <- sd(tr.y)
  L[i, 1] <- sum(L.FUN(tr.y, tr.p, s1))
  aic[i, 1] <- -2*L[i, 1] + d * 2
  bic[i, 1] <- -2*L[i, 1] + d * log(n)
  TE1[i, 1] <- mean((te.y - te.p)^2)

  s2 <- summary(fit.lm)$sigma
  L[i, 2] <- sum(L.FUN(tr.y, tr.p, s2))
  aic[i, 2] <- -2*L[i, 2] + d * 2
  bic[i, 2] <- -2*L[i, 2] + d * log(n)
  TE1[i, 2] <- mean((te.y - te.p)^2)
}

values <- rbind(apply(aic, 2, min), apply(bic, 2, min))
TEA <- TE1[apply(aic, 2, which.min)]
TEB <- TE1[apply(bic, 2, which.min)]
aic.pos <- grid[apply(aic, 2, which.min), ]
bic.pos <- grid[apply(bic, 2, which.min), ]
RES1 <- rbind(cbind(aic.pos, TE=TEA), cbind(bic.pos, TE=TEB))
rownames(RES1) <- c("M1", "M3", "M2", "M4")
res.Q61 <- RES1[order(rownames(RES1)), ]
```

M5 - M10

```
L <- aic <- bic <- TE2 <- matrix(0, length(lambda), 3)

fit.ls <- glmnet(tr.x, tr.y, alpha=1, lambda = lambda)
tr.p <- predict(fit.ls, as.matrix(tr.x))

for (i in 1:length(lambda)) {
  ds <- sum(coef(fit.ls)[,i] != 0)
  ##lm
  fit.lm <- lm(y ~ ., data=data.frame(y,x[,(fit.ls$beta[, i] != 0),drop=F]), subset=tran)
  te.p <- predict(fit.lm, data.frame(te.x))
  dm <- sum(fit.lm$coef != 0)

  ##M5 and M6
  si <- sd(tr.y)
  L[i, 1] <- sum(L.FUN(tr.y, tr.p[,i], si))
  aic[i, 1] <- -2 * L[i, 1] + ds * 2
  bic[i, 1] <- -2 * L[i, 1] + ds * log(n)
  TE2[i, 1] <- mean((te.y - te.p)^2)

  ##M7 and M8
  s4 <- summary(fit.lm)$sigma
  L[i, 2] <- sum(L.FUN(tr.y, tr.p[,i], s4))
  aic[i, 2] <- -2 * L[i, 2] + dm * 2
  bic[i, 2] <- -2 * L[i, 2] + dm * log(n)
  TE2[i, 2] <- mean((te.y - te.p)^2)

  ##M9 and M10
  s5 <- sqrt((1/(n - ds)) * sum((tr.y - tr.p[,i])^2) )
  L[i, 3] <- sum(L.FUN(tr.y, tr.p[,i], s5))
  aic[i, 3] <- -2 * L[i, 3] + ds * 2
  bic[i, 3] <- -2 * L[i, 3] + ds * log(n)
  TE2[i, 3] <- mean((te.y - te.p)^2)
}

maic <- apply(aic, 2, which.min)
mbic <- apply(bic, 2, which.min)
TEA <- TE2[apply(aic, 2, which.min)]
TEB <- TE2[apply(bic, 2, which.min)]

AIC.pos <- BIC.pos <- matrix(0, 3, 13)
colnames(AIC.pos) <- colnames(BIC.pos) <- colnames(x)

for (j in 1:length(maic)) {
  AIC.pos[j, ] <- ifelse(fit.ls$beta[, maic[j]] != 0, 1, 0)
  BIC.pos[j, ] <- ifelse(fit.ls$beta[, mbic[j]] != 0, 1, 0)
}

RES2 <- rbind(cbind(AIC.pos, TE=TEA), cbind(BIC.pos, TE=TEB))
rownames(RES2) <- c("M5", "M7", "M9", "M6", "M8", "M10")
res.Q62 <- rbind(RES2[order(rownames(RES2[1:5,])), ], M10=RES2[6,])
```

```
res.Q6 <- rbind(res.Q61, res.Q62)
res.Q6
```

```
##     zn indus chas nox rm age dis rad tax ptratio black lstat medv       TE
## M1   1     0    0   1  0   0   1   1   0       0     1     0    1 14.58761
## M2   0     0    0   0  0   0   0   1   0       0     0     1    0 13.68795
## M3   1     0    0   1  0   0   1   1   0       1     1     0    1 14.55877
## M4   0     0    0   0  0   0   0   1   0       0     1     1    0 14.50585
## M5   0     0    0   0  0   0   1   1   0       0     1     1    1 14.40786
## M6   0     0    0   0  0   0   0   1   0       0     1     1    1 14.44144
## M7   1     1    1   1  1   0   1   1   0       1     1     1    1 14.01744
## M8   0     0    0   0  0   0   1   1   0       0     1     1    1 14.40786
## M9   1     1    1   1  1   0   1   1   0       1     1     1    1 14.01744
## M10  0     0    0   0  0   0   1   1   0       0     1     1    1 14.40786
```