Homework Assignment 01

The Due Date : By 1:30pm, October, 11^{th} (Tuesday)
Your solution should include R codes and the answer of each question.
You need to upload your homework on `http://plato.pusan.ac.kr` for full credits.
You may collaborate on this problem but you must write up your own solution.

Open the data set `Boston` in the `R` package `MASS`. The data information is available with `?Boston`. It has a total of $n = 506$ observations with 14 variables, where the variable `crim` is considered as a response and the other 13 variables are considered as predictors. So, you can make the predictor `x` and the response `y` using the following `R` codes

```
> data(Boston)
> y <- Boston[, 1]
> x <- Boston[, -1]
```

We want to find the best subset among 13 predictors associated with the response `y`. In order to find the best model, you have to consider a total of $2^{13} - 1 = 8,191$ models. Let us define the log-likelihood function $l(\theta)$, AIC (Akaike information criterion) and BIC (Bayesian information criterion) as

$$l(\theta) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y_i - \beta_0 - x_i^{\mathrm{T}}\boldsymbol{\beta})^2}{2\sigma^2}\right)\right),$$

$$AIC(\theta) = -2l(\theta) + 2d, \qquad \text{and} \qquad BIC(\theta) = -2l(\theta) + d\log n,$$

respectively. The parameter vector $\theta = (\beta_0, \boldsymbol{\beta}, \sigma)$ and d is the number of regression coefficients in the model. Note that $d \geq 1$ since an intercept parameter β_0 is always in the model.

1. With the ordinary least square estimate $(\hat{\beta}_0^{ols}, \hat{\boldsymbol{\beta}}^{ols})$ and the plug-in estimate

$$\hat{\sigma}_1 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

   find the best subset models that can minimize AIC and BIC, respectively. For each best subset, clearly specify which variables are included in the final model along with the numerical values of AIC and BIC.

2. Repeat Q1, replacing $\hat{\sigma}_1$ with

$$\hat{\sigma}_2 = \sqrt{\frac{1}{n-d}\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0^{ols} - x_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{ols}\right)^2},$$

3. Fit lasso with all of 13 predictors to find the best subset among 1,000 different λ values. You don't need to consider 8,191 models here. Use the following `R` code to generate the λ values.

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

   With the lasso estimate $(\hat{\beta}_0^{lasso}, \hat{\boldsymbol{\beta}}^{lasso})$ and the plug-in estimate $\hat{\sigma}_1$, find the best subset models that can minimize AIC and BIC, respectively. For each best subset, clearly specify which variables are included in the final model along with the numerical values of AIC and BIC.

4. Repeat Q3, replacing $\hat{\sigma}_1$ with $\hat{\sigma}_2$

5. Repeat Q3, replacing $\hat{\sigma}_1$ with

$$\hat{\sigma}_3 = \sqrt{\frac{1}{n-d}\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0^{lasso} - x_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{lasso}\right)^2},$$

6. Randomly separate a training set (`tran`) and a test set (`test`), using the following `R` code

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

Find the best subset based on the training set, i.e., $n = 400$. In order to find the best subset, you should consider 5 different ways from Q1 to Q5. Since each question requires to find the best subset based on both AIC and BIC, you actually have 10 different ways to find the best subset. Let us denote them by $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_{10}$. Note that the best subsets from 10 different ways can be overlapped. After you find 10 best subsets from the training set, compute the test errors using

$$\frac{1}{106}\sum_{i=1}^{106}\left(y_i - \hat{\beta}_0^{ols} - x_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{ols}\right)^2$$

for each subset. For computation of the test error, $(\hat{\beta}_0^{ols}, \hat{\boldsymbol{\beta}}^{ols})$ should be estimated from the training set while $i = 1, \ldots, 106$ is an index set of the test set. Finally, provide the variable selection result from the training set and the test error (TE) for each subset using the following table,

	`zn`	`indus`	`chas`	$\cdots$	`lstat`	`medv`	TE
$\mathcal{M}_1$ (Q1/AIC)	0/1	0/1	0/1				
$\mathcal{M}_2$ (Q1/BIC)							
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$
$\mathcal{M}_9$ (Q5/AIC)							
$\mathcal{M}_{10}$ (Q5/BIC)				$\cdots$			

In the table, write '1' if the corresponding variable is included in the model, otherwise '0'. Who is winner?