REPORT

수강과목	:	빅데이터통계분석
담당교수	:	선호근
학 과	:	통계학과
학 번	:	201611531
이 름	:	정호재
제출일자	:	2020.10.20.

Homework Assignment 02

The Due Date : By Thursday, October, 20th in class
Your solution should include R codes and the answer of each question.
You need to upload your R codes 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 'Wage' in the R package 'ISLR'. The data information is available with ?Wage. The variable logwage is considered as a response variable and other 8 variables including 'year', 'age', 'maritl', 'race', 'education', 'jobclass', 'health', and 'health_ins' are considered as predictors. Since there are 8 predictors, a total of $2^8 - 1 = 255$ regression models should be considered.

```
> str(Wage)
'data.frame':    3000 obs. of  11 variables:
 $ year       : int  2006 2004 2003 2003 2005 2008 2009 2008 2006 2004 ...
 $ age        : int  18 24 45 43 50 54 44 30 41 52 ...
 $ maritl     : Factor w/ 5 levels "1. Never Married",..: 1 1 2 2 4 2 2 1 1 2 ...
 $ race       : Factor w/ 4 levels "1. White","2. Black",..: 1 1 1 3 1 1 4 3 2 1 ...
 $ education  : Factor w/ 5 levels "1. < HS Grad",..: 1 4 3 4 2 4 3 3 3 2 ...
 $ region     : Factor w/ 9 levels "1. New England",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ jobclass   : Factor w/ 2 levels "1. Industrial",..: 1 2 1 2 2 2 1 2 2 2 ...
 $ health     : Factor w/ 2 levels "1. <=Good","2. >=Very Good": 1 2 1 2 1 2 2 1 2 2 ...
 $ health_ins: Factor w/ 2 levels "1. Yes","2. No": 2 2 1 1 1 1 1 1 1 1 ...
 $ logwage    : num  4.32 4.26 4.88 5.04 4.32 ...
 $ wage       : num  75 70.5 131 154.7 75 ...
```

불필요한 변수들은 제거를 해준다.

```
> str(Wage)
'data.frame':    3000 obs. of  9 variables:
 $ year       : int  2006 2004 2003 2003 2005 2008 2009 2008 2006 2004 ...
 $ age        : int  18 24 45 43 50 54 44 30 41 52 ...
 $ maritl     : Factor w/ 5 levels "1. Never Married",..: 1 1 2 2 4 2 2 1 1 2 ...
 $ race       : Factor w/ 4 levels "1. White","2. Black",..: 1 1 1 3 1 1 4 3 2 1 ...
 $ education  : Factor w/ 5 levels "1. < HS Grad",..: 1 4 3 4 2 4 3 3 3 2 ...
 $ jobclass   : Factor w/ 2 levels "1. Industrial",..: 1 2 1 2 2 2 1 2 2 2 ...
 $ health     : Factor w/ 2 levels "1. <=Good","2. >=Very Good": 1 2 1 2 1 2 2 1 2 2 ...
 $ health_ins: Factor w/ 2 levels "1. Yes","2. No": 2 2 1 1 1 1 1 1 1 1 ...
 $ logwage    : num  4.32 4.26 4.88 5.04 4.32 ...
```

1. Let us define AIC (Akaike information criterion) and BIC (Bayesian information criterion) as

$$\text{AIC} = n\log\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2\right) + 2d, \quad \text{and} \quad \text{BIC} = n\log\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2\right) + d\log(n),$$

respectively. $\hat{f}(\cdot)$ is the estimated linear regression model and d is the number of regression coefficients in the model. For example, if 8 predictors are all included in the model, $d = 17$ due to dummy variables. Next, separate individuals into two groups based on their `logwage` value; group 1 is for $>$ `median(logwage)` and group 2 is for $\leq$ `median(logwage)`. Note that the sample size n should be regarded as a group size when AIC(BIC) is computed. For each group, find the best model among 255 models in terms of AIC and BIC, and provide the numerical values of AIC and BIC of the best models.

group1과 group2의 차원은 다음과 같다. (group1은 w1로 지정, group2는 w2로 지정)

```
> dim(w1)
[1] 1483    9
```

```
> dim(w2)
[1] 1517    9
```

```
> g.big1 <- regsubsets(logwage ~., data=w1, nvmax = 16)
> g.big2 <- regsubsets(logwage ~., data=w2, nvmax = 16)
```

각 모델의 적합모델을 다음과 같이 설정을 하고 각각의 AIC, BIC 값을 구한다.

적합모델에서의 which값을 살펴보면 8 개의 예측 변수가 모두 모형에 포함되어 있으면 더미 변수로 인해 d = 17으로 적용되었음을 알 수 있다. (절편 포함)
(regsubsets summary의 which : A logical matrix indicating which elements are in each model)

```
sg1 <- summary(g.big1)
> dim(sg1$which)
[1] 16 17
```

```
sg2 <- summary(g.big2)
> dim(sg2$which)
[1] 16 17
```

다음 표는 각 변수에 대한 degree별 PE값을 나타낸 데이터이다.

- Mallow's C_p statistics

$$C_p = \frac{1}{n}\left(\text{RSS} + 2d\hat{\sigma}^2\right)$$

where d is the total number of parameters used and $\hat{\sigma}^2$ is an estimate of the variance of the error ϵ.

- Akaike Information Criterion (BIC)

$$\text{AIC} = -2\log L + 2d$$

where L is the maximized value of the likelihood function for the estimated model. The AIC criterion is defined for a large class of models fit by maximum likelihood.

- For the linear model with normal errors, maximum likelihood and least squares are the same thing, so C_p and AIC are equivalent.

위의 모델은 선형회귀모델을 적합하므로 AIC와 Cp통계량은 동등한 값을 나타낸다.
따라서 내장된 함수 "cp", "bic"을 사용하여 AIC와 BIC를 구한다.

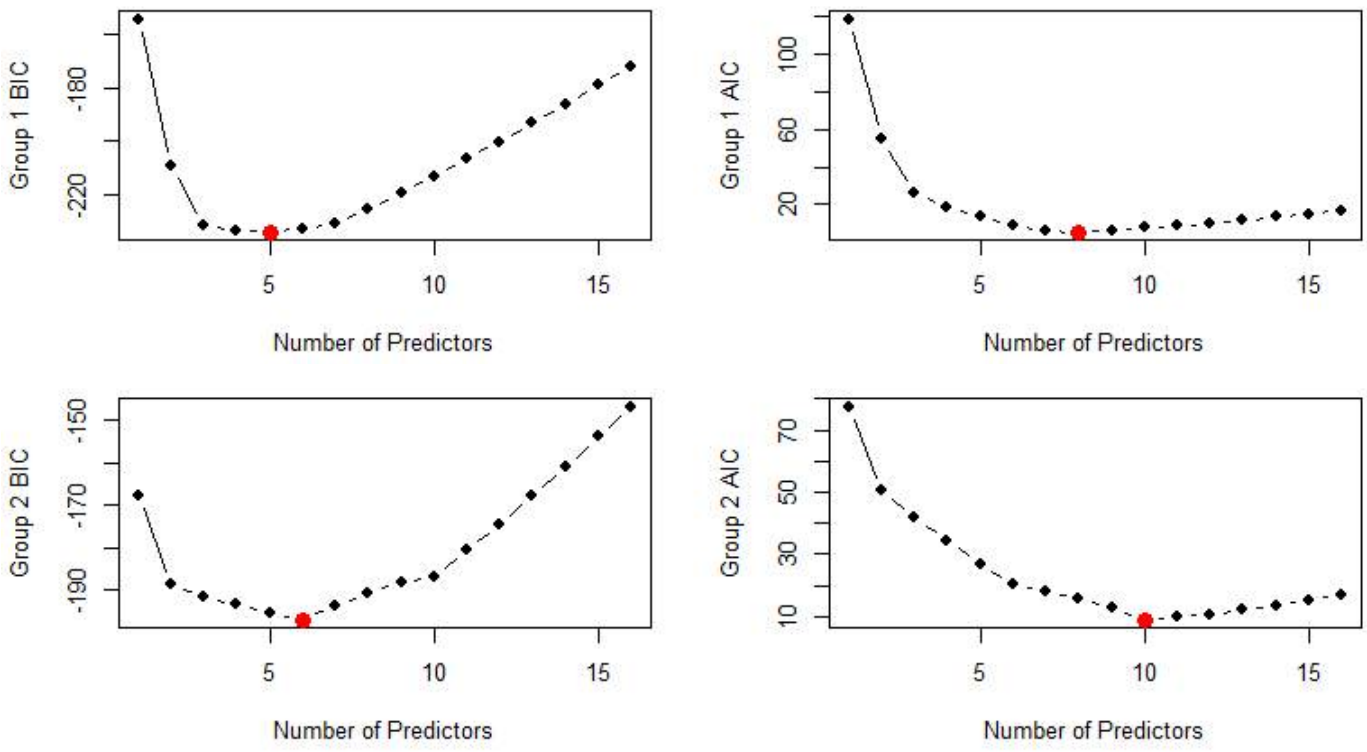
첫 번째 모델에 대한 BIC 및 Cp(AIC) 통계량은 다음과 같다.

```
> sg1.bic
        1         2         3         4         5         6         7         8         9
-154.4344 -209.2223 -231.4398 -233.6025 -233.8031 -232.5294 -230.8616 -225.5931 -219.4263
       10        11        12        13        14        15        16
-213.0813 -206.6511 -200.1949 -193.2488 -186.1830 -179.0721 -171.9182
> sg1.cp
         1          2          3          4          5          6          7          8
118.088586  54.602660  26.417340  18.865732  13.337497   9.316237   5.710667   5.692773
         9         10         11         12         13         14         15         16
  6.567687   7.619832   8.756879   9.920138  11.568318  13.334925  15.146192  17.000000
```

두 번째 모델에 대한 BIC 및 Cp(AIC) 통계량은 다음과 같다.

```
> sg2.bic
        1         2         3         4         5         6         7         8         9
-167.7365 -188.5640 -191.5840 -193.3798 -195.4158 -196.7936 -193.5626 -190.5848 -188.1268
       10        11        12        13        14        15        16
-186.7080 -180.5223 -174.3766 -167.8496 -160.9601 -153.8944 -146.7677
> sg2.cp
        1         2         3         4         5         6         7         8         9
77.528951 50.284640 41.674381 34.379685 26.897972 20.135431 18.030855 15.684653 12.833347
       10        11        12        13        14        15        16
 8.966961  9.838408 10.671072 11.881848 13.451543 15.195587 17.000000
```

이를 그래프로 나타내면 다음과 같다



각각의 그룹들의 최소인 AIC, BIC의 위치와 그때의 값은 다음과 같다.

```
> which.min(sg1$bic)
[1] 5
> sg1$bic[which.min(sg1$bic)]
[1] -233.8031
> which.min(sg1$cp)
[1] 8
> sg1$cp[which.min(sg1$cp)]
[1] 5.692773
```

```
> which.min(sg2$bic)
[1] 6
> sg2$bic[which.min(sg2$bic)]
[1] -196.7936
> which.min(sg2$cp)
[1] 10
> sg2$cp[which.min(sg2$cp)]
[1] 8.966961
```

각 그룹에 대해 선택된 AIC와 BIC 측면에서 가장 좋은 모델에서 선택된 계수들을 살펴보면 다음과 같다.

```
> RES
```

	w1_bic	w1_aic	w2_bic	w2_aic
(Intercept)	4.679319816	4.627327227	-12.180418727	-12.443245319
year	0.000000000	0.000000000	0.008231863	0.008340409
age	0.001545777	0.001504515	0.001600213	0.001450119
maritl2. Married	0.068159440	0.072944880	0.049464894	0.056944990
maritl3. Widowed	0.000000000	0.000000000	0.000000000	0.000000000
maritl4. Divorced	0.000000000	0.000000000	0.000000000	0.000000000
maritl5. Separated	0.000000000	0.000000000	0.000000000	0.090455830
race2. Black	0.000000000	0.000000000	0.000000000	0.000000000
race3. Asian	0.000000000	0.000000000	0.000000000	0.000000000
race4. Other	0.000000000	0.000000000	0.000000000	0.000000000
education2. HS Grad	0.000000000	0.046517808	0.000000000	0.047773527
education3. Some College	0.000000000	0.078674877	0.048092748	0.093985322
education4. College Grad	0.100721457	0.158661677	0.000000000	0.075109977
education5. Advanced Degree	0.208914251	0.266563748	0.000000000	0.083810535
jobclass2. Information	0.000000000	0.000000000	0.000000000	0.000000000
health2. >=Very Good	0.045847048	0.042608259	0.039886554	0.034480407
health_ins2. No	0.000000000	-0.034510973	-0.150456594	-0.145349769

2. Perform 10-fold cross validation to find the best model in terms of prediction error (PE), which can be defined as

$$\mathrm{PE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i - \hat{f}(x_i)\right)^2},$$

where $\hat{f}(\cdot)$ is the estimated linear regression using the training set and m is the size of the test set. Use the following R code to generate 10 folds.

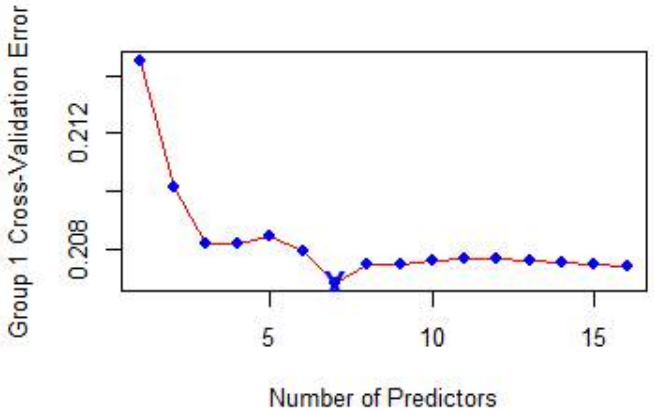
```
> RNGkind(sample.kind = "Rounding")
> set.seed(111)
> u1 <- sample(rep(seq(10), length=sum(Wage$logwage>median(Wage$logwage))))
> u2 <- sample(rep(seq(10), length=sum(Wage$logwage<=median(Wage$logwage))))
```

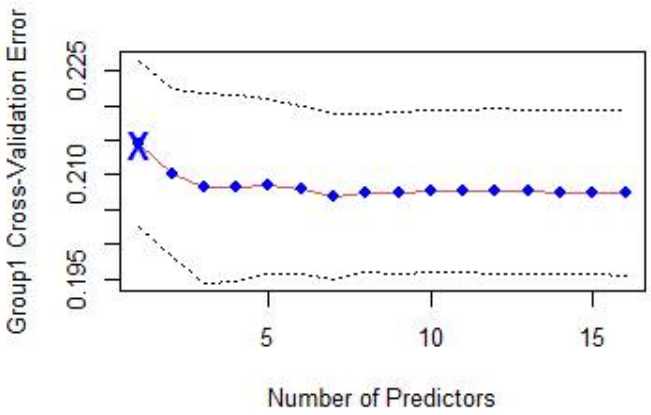
Note that u1 is for group 1 and u2 is for group 2. For each group, two best models should be determined based on the minimum PE (minPE) and one-standard-error rule (1SE), respectively. Therefore, your result of group 1 can be summarized as

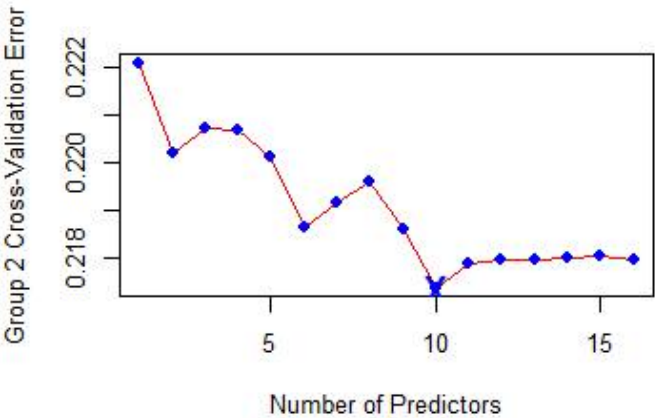
group 1	year	age	maritl	···	health	health_ins	PE
minPE	0 or 1						
1SE							

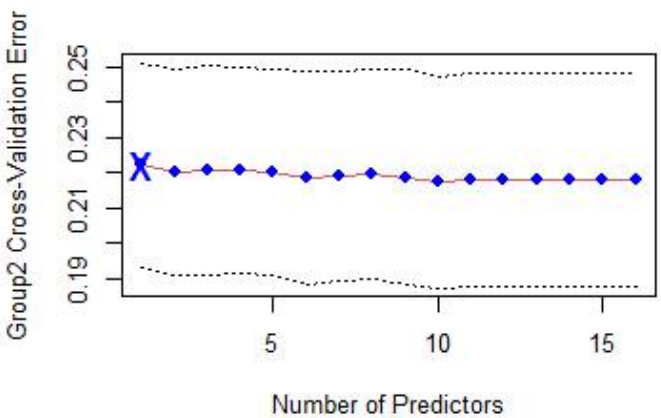
In the table, write '1' if the corresponding predictor is included in the model, otherwise '0'. In the one-standard-error rule, the best model has both the smallest model size (i.e., the smallest number of predictors) and the smallest PE among all models within one standard error boundary. You have to provide a table for group 2.

10-Fold에 대하여 그래프를 그리고 이때의 minPE와 1SE를 찾기 위해 그래프를 그려보았다.









이때 1se에서는 총 boundary의 교집합에서의 가장 작은 모델을 선택하였다.

각 그룹별로 계수는 다음과 같고

```
> RES
                                    w1_min      w1_1se          w2_min       w2_1se
(Intercept)                    4.667723075 4.8789263 -12.443245319   4.4640728
year                            0.000000000 0.0000000    0.008340409   0.0000000
age                             0.001510259 0.0000000    0.001450119   0.0000000
maritl2. Married                0.072714248 0.0000000    0.056944990   0.0000000
maritl3. Widowed                 0.000000000 0.0000000    0.000000000   0.0000000
maritl4. Divorced               0.000000000 0.0000000    0.000000000   0.0000000
maritl5. Separated              0.000000000 0.0000000    0.090455830   0.0000000
race2. Black                    0.000000000 0.0000000    0.000000000   0.0000000
race3. Asian                    0.000000000 0.0000000    0.000000000   0.0000000
race4. Other                    0.000000000 0.0000000    0.000000000   0.0000000
education2. HS Grad              0.000000000 0.0000000    0.047773527   0.0000000
education3. Some College        0.038276785 0.0000000    0.093985322   0.0000000
education4. College Grad        0.118257745 0.0000000    0.075109977   0.0000000
education5. Advanced Degree   0.226150041 0.1740528    0.083810535   0.0000000
jobclass2. Information          0.000000000 0.0000000    0.000000000   0.0000000
health2. >=Very Good             0.042666809 0.0000000    0.034480407   0.0000000
health_ins2. No                -0.035206033 0.0000000   -0.145349769 -0.1605577
```

해당 예측 변수가 모델에 포함되어 있으면 '1'을 쓰고 그렇지 않으면 '0'을 쓰면 다음과 같다.

```
> RES
                                     w1_min      w1_1se      w2_min      w2_1se
year                            0.0000000 0.0000000 1.0000000 0.0000000
age                             1.0000000 0.0000000 1.0000000 0.0000000
maritl2. Married                1.0000000 0.0000000 1.0000000 0.0000000
maritl3. Widowed                 0.0000000 0.0000000 0.0000000 0.0000000
maritl4. Divorced               0.0000000 0.0000000 0.0000000 0.0000000
maritl5. Separated              0.0000000 0.0000000 1.0000000 0.0000000
race2. Black                    0.0000000 0.0000000 0.0000000 0.0000000
race3. Asian                    0.0000000 0.0000000 0.0000000 0.0000000
race4. Other                    0.0000000 0.0000000 0.0000000 0.0000000
education2. HS Grad              0.0000000 0.0000000 1.0000000 0.0000000
education3. Some College        1.0000000 0.0000000 1.0000000 0.0000000
education4. College Grad        1.0000000 0.0000000 1.0000000 0.0000000
education5. Advanced Degree 1.0000000 1.0000000 1.0000000 0.0000000
jobclass2. Information          0.0000000 0.0000000 0.0000000 0.0000000
health2. >=Very Good             1.0000000 0.0000000 1.0000000 0.0000000
health_ins2. No                 1.0000000 0.0000000 1.0000000 1.0000000
PE                              0.2068295 0.2145292 0.2173833 0.2220984
```

Open the data set 'NCI60' in the R package 'ISLR'. The data information is available with ?NCI60. The gene expression data consists of 64 samples for 6,830 genes, where we assume that the first 50 genes are only relevant. That is, only 50 among 6,830 genes are associated with a response outcome. Type the following R codes to generate x and y with 300 lambda values and foldid for 5-fold cross validation.

```
> data(NCI60)
> x <- NCI60$data
> RNGkind(sample.kind = "Rounding")
> set.seed(123)
> beta <- rep(0, ncol(x))
> beta[1:50] <- runif(50, -2, 2)
> y <- x %*% beta + rnorm(nrow(x))
> lambda <- 10^seq(2, -2, length=300)
> foldid <- sample(rep(seq(5), length=length(y)))
```

For the 5-fold cross validation, you must use lambda and foldid to answer the following questions.

3. Apply the elastic-net for variable selection, where the tuning parameter α starts from 0 to 1 increased by 0.05. Perform the 5-fold cross validation to find out two optimal lambda values: $\hat{\lambda}_{\min}$ for the smallest prediction error and $\hat{\lambda}_{1se}$ for the one-standard-error rule. For computation of prediction errors, the default value of the glmnet package (mean squared errors) shoule be used. What is the optimal value of $\hat{\alpha}$ that minimizes the prediction error? Provide the numerical values of $\hat{\lambda}_{\min}$ and $\hat{\lambda}_{1se}$ for the corresponding $\hat{\alpha}$. How many variables are selected by $\hat{\lambda}_{\min}$ and $\hat{\lambda}_{1se}$, respectively?

NCI60 데이터의 차원은 다음과 같다.

```
> dim(x)
[1]   64 6830
```

5-Fold에서 glmnet package (mean squared errors)를 사용하여 alpha에 따른 값을 살펴보았다.
여기선 alpha=0.65일때 cvm이 가장 낮았다.

```
> cbind(alpha,cvm)
```

	alpha	cvm
[1,]	0.00	25.08736
[2,]	0.05	21.70247
[3,]	0.10	20.63715
[4,]	0.15	19.83034
[5,]	0.20	19.25049
[6,]	0.25	18.86278
[7,]	0.30	18.50379
[8,]	0.35	18.22518
[9,]	0.40	18.01368
[10,]	0.45	17.82816
[11,]	0.50	17.70375
[12,]	0.55	17.61360
[13,]	0.60	17.58490
[14,]	**0.65**	**17.57323**
[15,]	0.70	17.58874
[16,]	0.75	17.60875
[17,]	0.80	17.63605
[18,]	0.85	17.64097
[19,]	0.90	17.64707
[20,]	0.95	17.65626
[21,]	1.00	17.66994

위에서 구한 alpha 값을 적용하여 최적 alpha에 대한 λ^min 및 λ^1se의 숫자 값을 구해보자
> cvfit

Call: cv.glmnet(x = x, y = y, lambda = lambda, foldid = foldid, alpha = a)

Measure: Mean-Squared Error

    Lambda Measure   SE Nonzero
min 1.080   17.57 3.677   28
1se 2.481   21.22 2.713   12

λ^min = 1.080052 및 λ^1se = 2.481126으로 구해졌다.

> cvfit$lambda.min
[1] 1.080052

> cvfit$lambda.1se
[1] 2.481126

각 람다에 대한 회귀계수를 fit1, fit2로 지정하고 적용된 변수의 개수를 구하면
> fit1 <- coef(cvfit, s = "lambda.min")
> fit2 <- coef(cvfit, s = "lambda.1se")
> c(sum(as.matrix(fit1)!=0)-1, sum(as.matrix(fit2)!=0)-1)
[1] 28 12
λ^min 일 때 28개, λ^1se 일 때 12개로 구해졌다.(intercept를 제거하기 위하여 -1을 함)

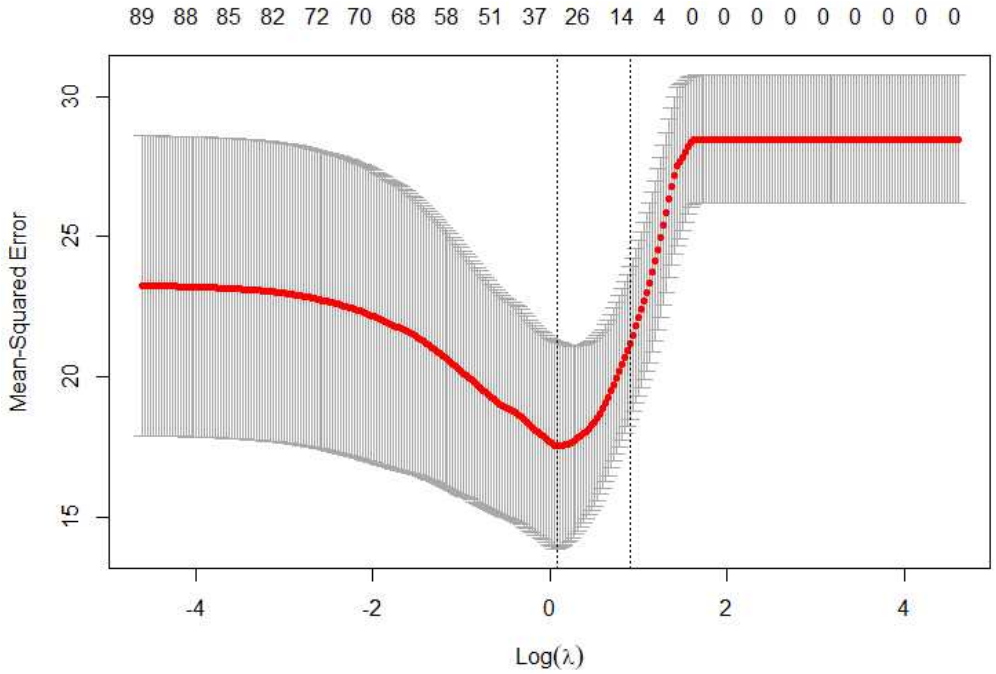
이때 적용된 변수명을 찾으면 다음과 같다.
> wh1 <- which(fit1!=0)
> w1 <- wh1[-1]-1

λ^min 일 때
> w1
 [1]   4   16   41   78 178 179 1096 1516 1577 1611 1612 2357 2486 2608 2665 3324 3385 3639
[19] 3841 3842 4503 4504 4536 4650 5020 5269 5907 6824

> wh2 <- which(fit2!=0)
> w2 <- wh2[-1]-1

λ^1se 일 때
> w2
 [1]   4   16   78 179 1096 1611 1612 2486 4027 4503 4504 4536

lambda값의 변화에 따른 MSE를 그래프로 나타내면 다음과 같다.



4. In the 5-fold cross validation, let us denote the number of training samples and the number of test samples by n_{tr} and n_{te}, respectively. Suppose that we select q variables from training samples, i.e., only $q+1$ regression coefficients including the intercept paramter are not zeros. Then, we newly define the prediction error as

$$\mathrm{PE}_{\mathrm{new}} = \sqrt{\frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \left(y_i - \hat{\beta}_0^{ls} - \sum_{j=1}^{q_0} x_{ij} \hat{\beta}_j^{ls} \right)^2},$$

where $q_0 = \min(q, n_{tr} - 1)$ and $\hat{\beta}_j^{ls}$ is the ordinary least square estimate for the jth selected variable. For each λ value, first find q variables that have nonzero regression coefficients and then compute the ordinary least square estimate for the correponding q variables. Note that the ordinary least square esitmate cannot be computed when $q \geq n_{tr}$, so we only select the first $n_{tr} - 1$ variables in this case. Your optimal $\hat{\lambda}_{\mathrm{min}}$ and $\hat{\lambda}_{\mathrm{1se}}$ should be determined based on $\mathrm{PE}_{\mathrm{new}}$ of the 5-fold cross validation. For lasso ($\alpha = 1$), find the numerical values of $\hat{\lambda}_{\mathrm{min}}$ and $\hat{\lambda}_{\mathrm{1se}}$, and provide the number of selected variables for the corresponding $\hat{\lambda}_{\mathrm{min}}$ and $\hat{\lambda}_{\mathrm{1se}}$, respectively.

5-Fold CV에서 새로운 PE값을 적용시켜 최적의 lambda값을 구하고자 한다. 이때 q ≥ ntr 경우는 이 경우 첫 번째 ntr − 1 변수 만 선택하였다.
라쏘 (α = 1) 의 경우에서 결정된 최적의 λ^min는 178번째의 0.428654140842093이고
최적의 λ^1se는 176번째의 0.4558929이었다.

```
> which.min(PE[,2])
0.428654140842093
              178
> up <- which(PE[,2] < PE[which.min(PE[,2]),3]) # 최솟값 PE 범위에 있는 모델
> min(up)
[1] 176
> lambda[min(up)]
[1] 0.4558929
```

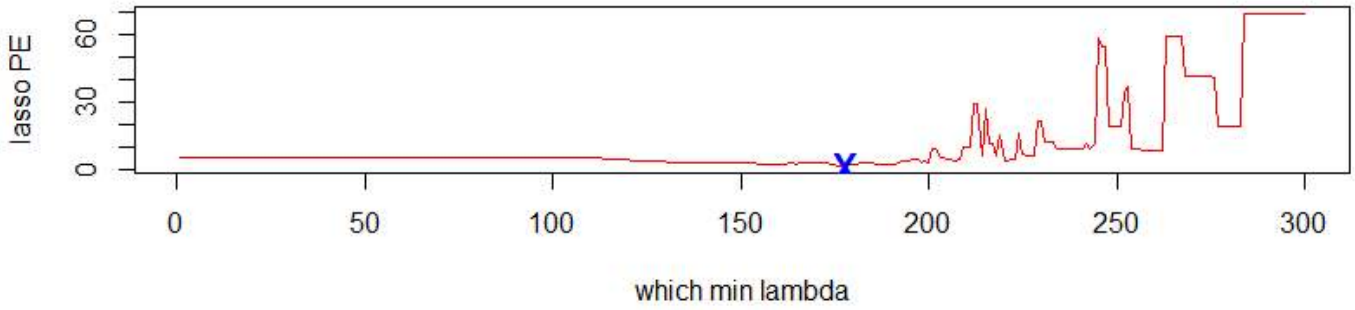
라쏘모델에서 최적의 λ^min값인 0.428654140842093에서 선택된 변수는 다음과 같고 이때 변수의 개수는 39개이다.

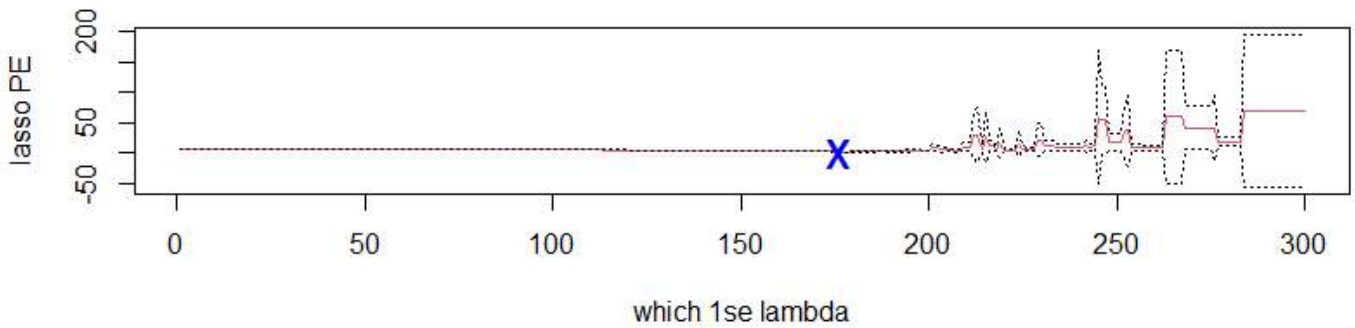
```
> min <- coef(g, lambda[which.min(PE[,2])])
> which(min!=0)-1  # first term means intercept
 [1]    0    4   16   78  135  179  881 1073 1215 1516 1577 1611 2346 2357 2608 2632 2665 3324
[19] 3367 3385 3399 3639 3841 3854 4503 4504 4536 4633 4650 5020 5183 5223 5258 5269 6269 6308
[37] 6401 6424 6785 6824
> dim(x[,which(min!=0)[-1]-1])
[1] 64 39
```

라쏘모델에서 최적의 λ^1se값인 0.4558929에서 선택된 변수는 다음과 같고 이때 변수의 개수는 39개이다.
이 때 0은 절편이다.

```
> one_se <- coef(g, lambda[min(up)])
> which(one_se!=0)-1   # first term means intercept
 [1]    0    4   16   78  135  179  881 1073 1096 1215 1516 1577 1611 2346 2357 2608 2632 2665
[19] 3324 3367 3385 3399 3639 3841 3854 4503 4504 4536 4650 5020 5183 5223 5258 5269 6269 6308
[37] 6401 6424 6785 6824
> dim(x[,which(one_se!=0)[-1]-1])
[1] 64 39
```

다음은 라쏘모형을 lambda의 위치에 따른 PE값을 그래프로 나타낸 것이다.





5. Repeat Q4 with the elastic-net, where the tuning parameter α starts from 0 to 1 increased by 0.05. Suppose that $\hat{\alpha}$ minimizes $\mathrm{PE}_{\mathrm{new}}$ among all α. Find the optimal value of $\hat{\alpha}$, and the corresponding $\hat{\lambda}_{\mathrm{min}}$ and $\hat{\lambda}_{\mathrm{1se}}$. Also provide the number of selected variables for $\hat{\lambda}_{\mathrm{min}}$ and $\hat{\lambda}_{\mathrm{1se}}$, respectively.

α는 0에서 1로 시작하여 0.05만큼 증가하는 것으로 설정하여 elastic-net으로 Q4를 반복하였다.
5-Fold CV를 사용하여 각 alpha마다 300개의 lambda를 설정하여 5개의 폴드에 대한 각각의 PE값을 추출하였다.
그리고 각 alpha의 lambda별 5개 폴드에 대한 PE값을 평균내었다.
그 뒤 각 alpha의 lambda별 300개의 PE 값 중에서 (lambda가 300개) 최솟값을 추출하여 다음과 같이 나타내었다.

```
> cbind(alpha,PE)
      alpha        PE
 [1,]  0.00 31.302530
 [2,]  0.05  3.128385
 [3,]  0.10  3.129302
 [4,]  0.15  3.026073
 [5,]  0.20  3.026073
 [6,]  0.25  3.026073
 [7,]  0.30  3.026073
 [8,]  0.35  3.030896
 [9,]  0.40  3.030896
[10,]  0.45  3.030896
[11,]  0.50  2.924144
[12,]  0.55  2.889461
[13,]  0.60  2.714846
[14,]  0.65  2.560464
[15,]  0.70  2.560464
[16,]  0.75  1.752796
[17,]  0.80  1.752796
[18,]  0.85  1.765382
[19,]  0.90  1.959588
[20,]  0.95  1.530152
[21,]  1.00  1.340013
```

최적의 α는 1일 때이다.

```
> a <- alpha[which.min(PE)]
> a
[1] 1
```

α = 1 으로 설정하여 Q4를 다시 반복하면

최적의 alpha에서 결정된 최적의 λ^min는 178번째의 0.428654140842093이고
최적의 λ^1se는 176번째의 0.4558929이었다.

```
> which.min(PE[,2])
0.428654140842093
              178
> up <- which(PE[,2] < PE[which.min(PE[,2]),3]) # 최솟값 PE 범위에 있는 모델
> min(up)
[1] 176
> lambda[min(up)]
[1] 0.4558929
```

최적의 alpha에서 적합된 모델에서 최적의 λˆmin값인 0.428654140842093에서 선택된 변수는 다음과 같고 이때 변수의 개수는 39개이다. (0은 절편이다.)

```
> min <- coef(g, lambda[which.min(PE[,2])])
> which(min!=0)-1  # first term means intercept
 [1]    0    4   16   78  135  179  881 1073 1215 1516 1577 1611 2346 2357 2608 2632 2665 3324
[19] 3367 3385 3399 3639 3841 3854 4503 4504 4536 4633 4650 5020 5183 5223 5258 5269 6269 6308
[37] 6401 6424 6785 6824
> dim(x[,which(min!=0)[-1]-1])
[1] 64 39
```

최적의 alpha에서 적합된 모델에서 최적의 λˆ1se값인 0.4558929에서 선택된 변수는 다음과 같고 이때 변수의 개수는 39개이다.

```
> one_se <- coef(g, lambda[min(up)])
> which(one_se!=0)-1  # first term means intercept
 [1]    0    4   16   78  135  179  881 1073 1096 1215 1516 1577 1611 2346 2357 2608 2632 2665
[19] 3324 3367 3385 3399 3639 3841 3854 4503 4504 4536 4650 5020 5183 5223 5258 5269 6269 6308
[37] 6401 6424 6785 6824
> dim(x[,which(one_se!=0)[-1]-1])
[1] 64 39
```

다음은 라쏘모형을 lambda의 위치에 따른 PE값을 그래프로 나타낸 것이다.