1. Description

▶ 회귀 모델에서 어떤 변수를 어떻게 선택하느냐는 중요한 문제이다. 모델을 비교하는 척도에는 MSE 를 바탕으로 한 Cp, AIC, BIC, adjusted R^2가 있다. 먼저, ***Cp*** 는 RSS 와 변수의 개수를 가지고 MSE 를 추정하는 방식이다. 변수의 개수에 패널티를 주어, 개수가 많아지면 MSE 가 증가한다. ***AIC*** 는 Cp 를 scaling 한 척도이다. ***BIC*** 는 Cp, AIC 와 비슷하지만 전체 관측치 수에 log 를 취하여 대부분의 경우 BIC 가 Cp 보다 큰 값을 갖는다. ***adjusted R^2***는 자유도와 변수의 개수가 반영된 척도로 R^2이 과적합이 되는 경향이 발견되기 힘든 점을 보완한 값이다.

▶ best subset selection 은 가능한 모든 모델의 정확성을 계산 및 비교하여 "가장 좋은" 모델을 선정하는 방법이다. 그 중 ***forward selection*** 은 모델에 변수를 하나씩 추가해 나가면서 best subset 을 찾아가는 방법이다. 변수가 모델에 추가되는 순서가 중요하고 진짜 최적 모델을 놓치는 경우가 발생할 수 있다. ***backward selection*** 은 모델에 변수를 하나씩 제거하면서 찾아가는 방법이다. 두 방법의 단점을 보완하여, 변수를 추가하거나 제거하는 방법을 ***stepwise selection*** 이라고 한다.

▶ OLS 에서 관측치보다 변수가 많아질수록 beta 의 분산이 무한에 수렴하게 되는 것을 보완하는 모델이 Ridge 와 Lasso 이다. ***Ridge*** 는 일반적인 OLS 에서 RSS 를 최소화하는 식에서 계수에 대한 패널티를 $\lambda\sum_{j=1}^{n}\beta_j^2$로 주고, ***Lasso*** 는 $\lambda\sum_{j=1}^{n}|\beta_j|$로 추정한다. λ가 0 에 근사할수록 bias 는 줄고 분산이 높아지는 성질을 갖게 되며, 반대로 무한대에 근사할수록 bias, 낮은 분산 값을 갖게 된다. 두 모델의 다른 점은 Ridge 를 통한 계수는 0 에 근사하지만 0 이 되지는 않으나, lasso 는 0 으로 보내 변수선택을 가능하게 한다.

▶ ***PCR*** (Principal Component Regression)은 독립 변수들의 주성문을 추출한 후, 주성분들을 이용해서 회귀 모델을 만드는 방법이다. Lasso 처럼 regularization 효과를 줄 수 있어 과적합을 완화시킬 수 있지만 실제 독립변수들의 전체 영향력을 부분적으로 반영한 변수들이기 때문에 각 조건의 영향력을 파악하기가 거의 불가능해지는 문제가 있다. PLS 는 PCR 과 기본 개념은 비슷하지만 변수들의 변환 방식이 다르다. PCR 은 독립 변수의 분산을 최대로 하는 축을 찾아 데이터를 전사하는 방식으로 독립변수만 변형하지만, ***PLS*** (Partial Least Square)은 종속 변수와 독립 변수의 관계를 가장 잘 설명하는 축을 찾아 전사하는 방식으로 종속 변수, 독립 변수 모두를 변형한다.

2. Implement

Lab : Linear Regression

- Hitters 데이터를 이용해서 야구 선수들의 연봉 예측하는 문제를 best subset selection 문제에 적용해보았다. 먼저, N 개의 설명 변수가 있을 때, 각 변수를 추가하거나 뺀 총 2N 개의 회귀 모델을 만들고 이들 모두를 비교해보는 함수인 regsubsets()를 사용하였다. nvmax option 을 통해 사용할 최대의 변수 개수를 정하고 method 로 전진선택법이면 "forward", 후진제거법이면 "backward"로 지정할 수 있다.

- summary(model)을 이용해서 R^2, RSS, C_p, BIC 를 통해 최적을 모델은 선택하였다. R^2가 클수록, C_p, BIC 는 작을수록 모델이 좋다고 할 수 있다.

- glmnet()은 ridge 와 lasso regression 을 적용할 수 있는 함수이다. lm(y~x)과는 다르게 matrix 형태로 glmnet(x,y)로 표현하며, alpha=0 이면 ridge, alpha=1 이면 lasso 이다. lambda 는 grid search 를 이용하여, MSE 를 최소로 하는 값을 찾는다. Cross-validation 을 할 수 있는 cv.glmnet() 함수를 이용해 lambda 를 구할 수도 있다.

- PCR 은 pls library 의 pcr()를 이용하고 scale=TRUE 로 각 변수를 표준화 시키거나 validation="CV"를 통해 교차검정도 할 수 있다. PLS 는 plsr() 함수로 적용시키고 사용법은 pcr()과 같다.

5. It is well-known that ridge regression tends to give similar coefficient values to correlated variables, whereas

the lasso may give quite different coefficient values to correlated variables. We will now explore this property in a very simple setting. Suppose that *n* = 2, *p* = 2, *x*11 = *x*12, *x*21 = *x*22. Furthermore, suppose that *y*1+*y*2 = 0 and *x*11+*x*21 = 0 and *x*12+*x*22 = 0, so that the estimate for the intercept in a least squares, ridge regression, or lasso model is zero: $\widehat{\beta 0}$ = 0.

(a) Write out the ridge regression optimization problem in this setting.

$(y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{12})^2 + (y_2 - \hat{\beta}_1 x_{21} - \hat{\beta}_2 x_{22})^2 + \lambda(\hat{\beta}_1^2 + \hat{\beta}_2^2)$ 를 최소화하는 것이 목적인 모델이 ridge이다.

(b) Argue that in this setting, the ridge coefficient estimates satisfy $\widehat{\beta 1}$ = $\widehat{\beta 2}$.

(a)의 식을 전개하면 다음과 같다.

$$(y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{12})^2 + (y_2 - \hat{\beta}_1 x_{21} - \hat{\beta}_2 x_{22})^2 + \lambda(\hat{\beta}_1^2 + \hat{\beta}_2^2)$$
$$= (y_1^2 + \hat{\beta}_1^2 x_{11}^2 + \hat{\beta}_2^2 x_{12}^2 - 2\hat{\beta}_1 x_{11} y_1 - 2\hat{\beta}_2 x_{12} y_1 + 2\hat{\beta}_1 \hat{\beta}_2 x_{11} x_{12})$$
$$+ (y_2^2 + \hat{\beta}_1^2 x_{21}^2 + \hat{\beta}_2^2 x_{22}^2 - 2\hat{\beta}_1 x_{21} y_2 - 2\hat{\beta}_2 x_{22} y_2 + 2\hat{\beta}_1 \hat{\beta}_2 x_{21} x_{22})$$
$$+ \lambda\hat{\beta}_1^2 + \lambda\hat{\beta}_2^2$$

양변을 $\widehat{\beta 1}$ 에 대해 미분한다. $\frac{\partial}{\partial \hat{\beta}_1} : (2\hat{\beta}_1 x_{11}^2 - 2x_{11} y_1 + 2\hat{\beta}_2 x_{11} x_{12}) + (2\hat{\beta}_1 x_{21}^2 - 2x_{21} y_2 + 2\hat{\beta}_2 x_{21} x_{22}) + 2\lambda\hat{\beta}_1 = 0$

$x_{11} = x_{12} = x_1$, $x_{21} = x_{22} = x_2$ 라고 하면

$$(\hat{\beta}_1 x_1^2 - x_1 y_1 + \hat{\beta}_2 x_1^2) + (\hat{\beta}_1 x_2^2 - x_2 y_2 + \hat{\beta}_2 x_2^2) + \lambda\hat{\beta}_1 = 0$$
$$\hat{\beta}_1(x_1^2 + x_2^2) + \hat{\beta}_2(x_1^2 + x_2^2) + \lambda\hat{\beta}_1 = x_1 y_1 + x_2 y_2$$

이고,

$x_1 + x_2 = 0$ 이기 때문에, $\lambda\hat{\beta}_1 = x_1 y_1 + x_2 y_2 + 2\hat{\beta}_1 x_1 x_2 + 2\hat{\beta}_2 x_1 x_2$ 이다.

같은 방법으로, $\widehat{\beta 2}$ 역시

$$\lambda\hat{\beta}_2 = x_1 y_1 + x_2 y_2 + 2\hat{\beta}_1 x_1 x_2 + 2\hat{\beta}_2 x_1 x_2$$ 이므로

$$\lambda\hat{\beta}_1 = \lambda\hat{\beta}_2 \quad \rightarrow \quad \hat{\beta}_1 = \hat{\beta}_2$$

(c) Write out the lasso optimization problem in this setting.

$(y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{12})^2 + (y_2 - \hat{\beta}_1 x_{21} - \hat{\beta}_2 x_{22})^2 + \lambda(|\hat{\beta}_1| + |\hat{\beta}_2|)$ 를 최소화하는 것이 목적인 모델이 lasso이다.

(d) Argue that in this setting, the lasso coefficients $\widehat{\beta 1}$ and $\widehat{\beta 2}$ are not unique—in other words, there are many possible solutions to the optimization problem in (c). Describe these solutions.

비슷한 방법으로 편미분하면, $\frac{\partial}{\partial \hat{\beta}}(\lambda|\beta|) : \lambda\frac{|\beta|}{\beta}$ 이고 $\lambda\frac{|\beta_1|}{\beta_1} = \lambda\frac{|\beta_2|}{\beta_2}$ 이므로 해가 unique 하지 않고 둘다

양수이거나 둘다 음수인 값이다.

9. We will predict the number of applications received using the other variables in the College data set.

(a) Split the data set into a training set and a test set.

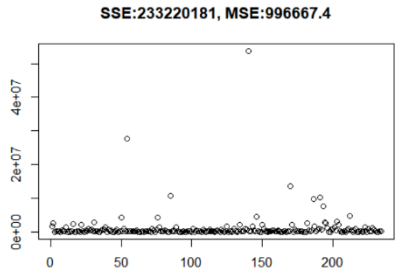
College 데이터를 train:test=7:3으로 나누었고 train의 차원은 (543,18), test는 (234,18)이다.

(b) Fit a linear model using least squares on the training set, and report the test error obtained.

```
Call : lm(formula = Apps ~ ., data = train)
Coefficients:
(Intercept)   PrivateYes       Accept       Enroll
 -164.88351   -602.60121      1.28244     -0.37473
  Top10perc    Top25perc  F.Undergrad  P.Undergrad
   43.42687    -11.44400      0.08491      0.05034
   Outstate   Room.Board        Books     Personal
   -0.03897      0.22259     -0.02608     -0.05946
        PhD     Terminal   S.F.Ratio  perc.alumni
   -6.39782     -2.97670     -9.70306     -8.02855
     Expend    Grad.Rate
    0.03298      8.68729
```

Train data를 OLS에 적용했을 때, 18개 변수에 대한 회귀계수는 다음과 같다.

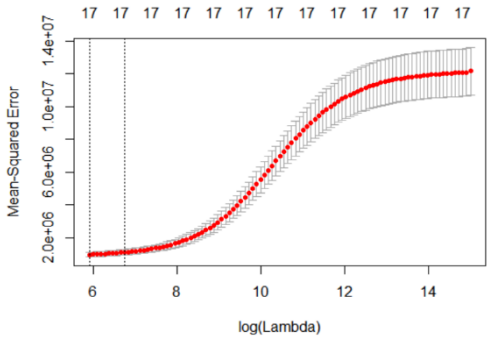
만들어진 모델에 test data을 넣고 예측값과 실제값의 차인 오차 제곱의 분포를 살펴보았다.



(c) Fit a ridge regression model on the training set, with λ chosen by cross-validation. Report the test error obtained.

```
(Intercept)  PrivateYes     Accept     Enroll
-2481.652    -590.8609     0.5552962  0.7436282
Top10perc  Top25perc  F.Undergrad  P.Undergrad
16.43246   9.175878    0.1255869    0.09598456
Outstate     Room.Board   Books      Personal
0.01251    0.19529        0.28574    0.029198
PhD        Terminal     S.F.Ratio      perc.alumni
3.2201     0.952572      17.32374      -11.21212
Expend     Grad.Rate
0.05378    11.1161
```

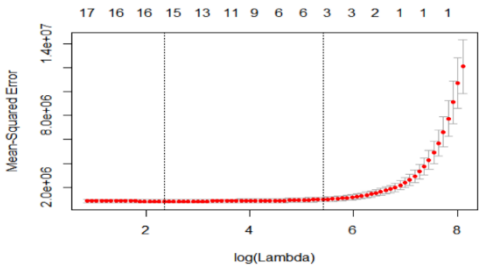
MSE를 최소화하는 λ 는 363.23으로
그때의 MSE는 379705이다.



(d) Fit a lasso model on the training set, with λ chosen by cross validation. Report the test error obtained, along with the number of non-zero coefficient estimates.

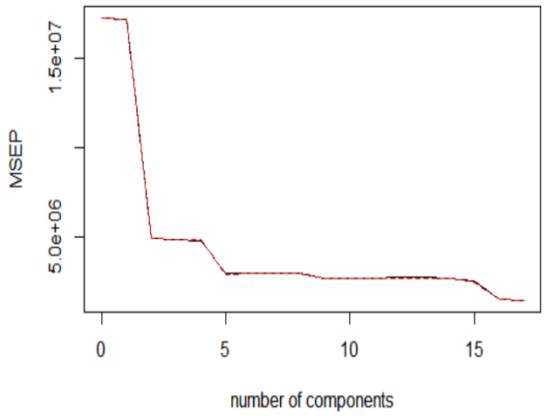
```
(Intercept)  PrivateYes   Accept   Enroll
 -429.5279   -501.5649    1.5723   -0.5168
Top10perc   Top25perc  F.Undergrad P.Undergrad
 42.2696     -8.2746         .       -0.0007
Outstate    Room.Board    Books     Personal
-0.0713       0.1654        .        0.0388
  PhD     Terminal   S.F.Ratio    perc.alumni
-4.3488  -6.4598     7.5950          -1.4322
Expend   Grad.Rate
0.0668     6.1402
```

MSE를 최소화하는 λ 는 3.6689으로
그때의 MSE는 992521.76이다.
회귀계수가 0인 변수는 F.Undergrad와 Books로 변수 18개 중에서 유의한 변수는 16개이다.



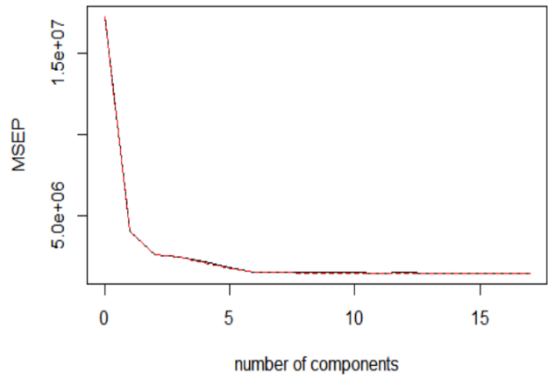
(e) Fit a PCR model on the training set, with *M* chosen by cross validation. Report the test error obtained, along

with the value of *M* selected by cross-validation.



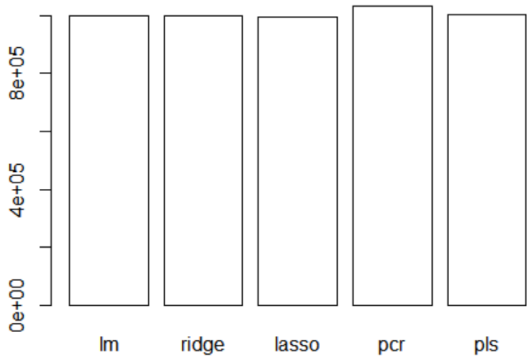
함수 pcr()을 사용해서 cv를 적용시켜 모델을 fitting 시켰다. cross-validation별로 MSE에 대한 그래프는 다음과 같다.
그래프를 보면 M=16일 때 MSE가 최소인 것을 볼 수 있고, 이때의 MSE는 1031142이다.

(f) Fit a PLS model on the training set, with *M* chosen by cross validation. Report the test error obtained, along with the value of *M* selected by cross-validation.



함수 plsr()을 사용해서 cv를 적용시켜 모델을 fitting 시켰다. cross-validation별로 MSE에 대한 그래프는 다음과 같다.
그래프를 보면 M=10일 때 MSE가 최소인 것을 볼 수 있고, 이때의 MSE는 1002529이다.

(g) Comment on the results obtained. How accurately can we predict the number of college applications received? Is there much difference among the test errors resulting from these five approaches?



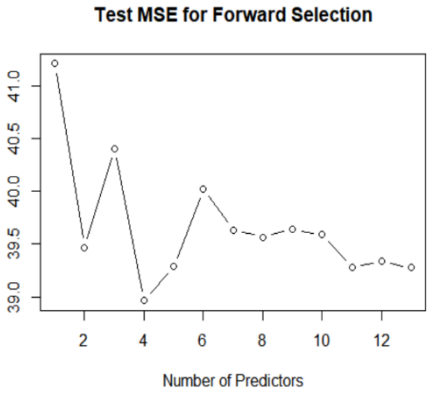
lm	Ridge	Lasso	PCR	PLS
996667.4	999514.5	992521.6	1031142.3	1002529.5

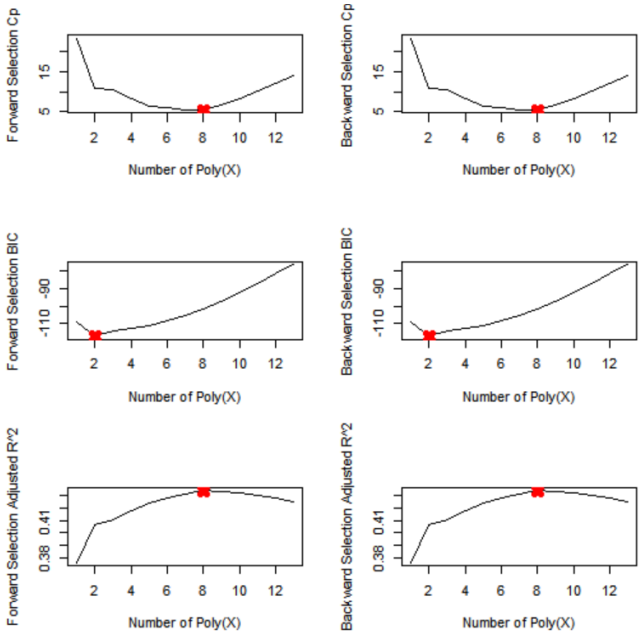
다음은 다섯가지 모델에 대한 MSE값이다.
Lasso, Ridge, lm, PLS, PCR 순으로 예측률이 낮았다.

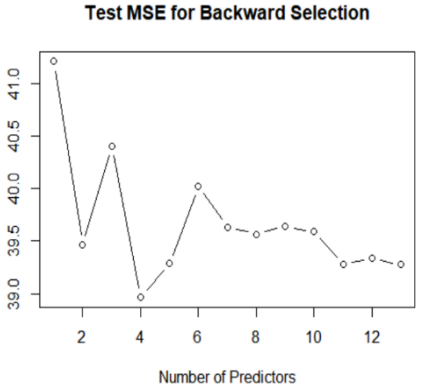
11. We will now try to predict per capita crime rate in the Boston data set.

(a) Try out some of the regression methods explored in this chapter, such as best subset selection, the lasso, ridge regression, and PCR. Present and discuss results for the approaches that you consider.

① forward/backward selection



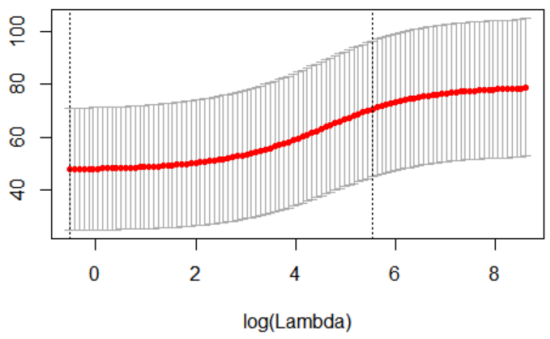




lm()으로 모델을 만들고 step(method="forward","backward") 함수를 써서 best subset selection을 하였다. forward와 backward의 결과가 같고, 총 13개 변수 중에서 채택된 변수는 4개이다.

② Ridge

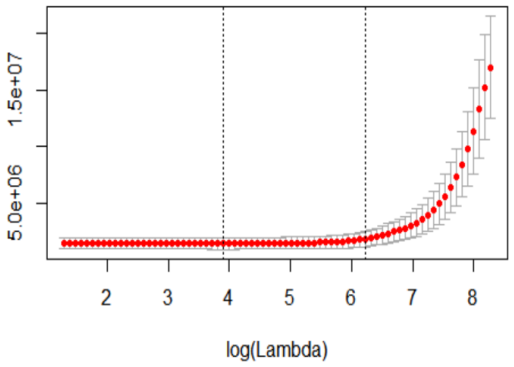
(Intercept)	4.9521	age	0.0056
zn	0.0371	dis	-0.9086
indus	-0.0614	tax	0.0029
chas	-0.7730	rad	0.4399
nox	-8.4613	lstat	0.1933
rm	1.2006	black	-0.0039
ptratio	-0.1817	medv	-0.1759



Cross-validation을 통해서 MSE를 최소로 하는 lambda는 0.8679이다.
그때의 ridge 모델에서 회귀계수는 다음과 같다. nox의 계수는 다른 값에 비해 굉장히 작은 음수고 rm은 비교적 큰 양수의 계수를 갖고 있으므로 crime을 예측할 때 중요한 변수라고 할 수 있다.

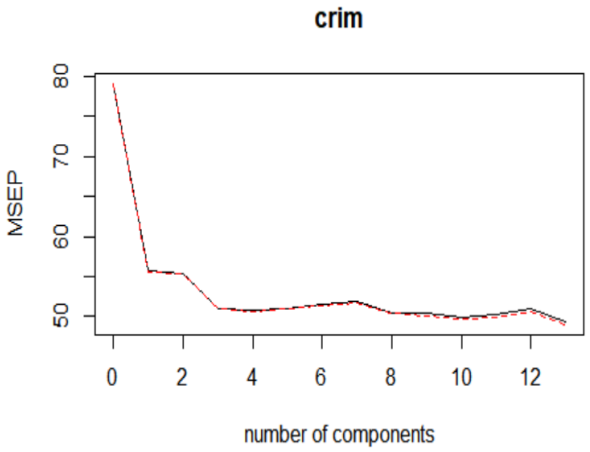
③ Lasso

(Intercept)	7.2046	tax	.
zn	0.03708	dis	-0.9716
indus	-0.0340	rad	0.5170
chas	-0.5188	ptratio	-0.2205
nox	-9.6638	lstat	0.1766
rm	1.1871	medv	-0.1931
age	.	black	-0.0024



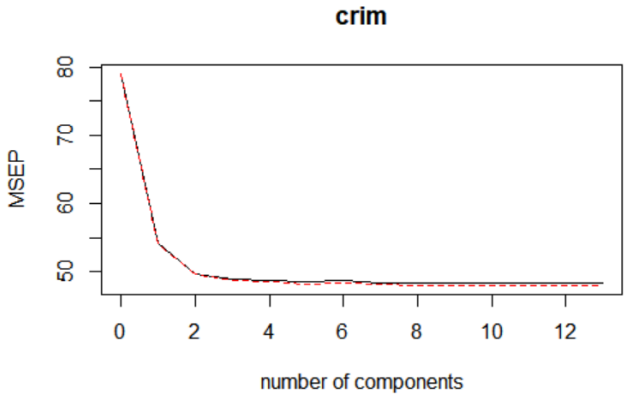
Cross-validation을 통해서 MSE를 최소로 하는 lambda는 0.0828이다.
그때의 lasso 모델에서 회귀계수는 다음과 같다. nox의 계수는 다른 값에 비해 굉장히 작은 음수이므로 crime을 예측할 때 중요한 변수라고 할 수 있다. Lasso는 계수를 0으로 보내는 특징이 있고, Boston 데이터에서는 age와 tax가 의미 없는 변수로 나왔다.

④ PCR



함수 pcr()을 사용해서 cv를 적용시켜 모델을 fitting 시켰다.
cross-validation별로 MSE에 대한 그래프는 다음과 같다. 그래프를 보면 M=10일 때 MSE가 최소인 것을 볼 수 있다.

⑤ PCA



함수 plsr()을 사용해서 cv를 적용시켜 모델을 fitting 시켰다.
cross-validation별로 MSE에 대한 그래프는 다음과 같다. 그래프를 보면 M=3일 때 MSE가 최소인 것을 볼 수 있다.

(b) Propose a model (or set of models) that seem to perform well on this data set, and justify your answer. Make sure that you are evaluating model performance using validation set error, cross validation, or some other reasonable alternative, as opposed to using training error.

lm	Ridge	Lasso	PCR	PLS
38.96427	38.3787	38.3766	39.6653	39.2045

5가지 방법에 대한 validation set error는 다음과 같다. 5가지 방법의 MSE는 별 차이 없지만 Lasso, Ridge, lm, PLS, PCR 순으로 좋다. 따라서 Boston data에서 선택된 subset 모델은 Lasso이다.

(c) Does your chosen model involve all of the features in the data set? Why or why not?

Ridge와 거의 차이는 없지만 변수가 줄어서 MSE가 낮아져 예측력이 높아졌기 때문에 age, tax를 제외하고 11개의 변수를 포함한 lasso 모형을 최종 모형으로 선택했다.

3. Discussion

▶ 이번 장에서는 회귀모델에서 최적의 변수 조합을 찾는 법에 대해서 배웠다.

▶ Ridge 와 Lasso 는 덜 flexible 하지만, 분산을 줄이면서 bias 를 늘리므로 예측력을 증가시킨다. 그에 반해 비선형 모델은 좀더 flexible 하고 bias 가 낮지만 강한 분산을 갖고 있는 모델이다.

▶ 6.5 에서는 Ridge 와 Lasso 회귀식이 최적화 되는 베타들의 값을 구하는 것을 수식을 통해 증명해 보았다.

▶ 6.9 와 6.11 에서는 College, Boston data 에 대해서 train, test 로 나누고 cross-validation 을 통해 최적의 모델을 만들었다. Ridge 와 Lasso 는 MSE 를 최소화 시키는 lambda 를 구했고, PCR 과 PLS 는 MSE 를 최소화 시키는 M 을 구했다. 두 데이터 모두 모델 별로 큰 차이는 없었지만 Lasso, Ridge, lm, PLS, PCR 순으로 예측력이 좋았다.