8 장 Penalized Regression

CONTENTS

8.1 서론

- 앞서 5.2절에서 다중공선성에 대한 해결책으로 벌점회귀(능형, LASSO, Elestic Net)를 언급한 바 있다. 다중공선성이 심한 경우(예측변수들 간의 상관이 매우 높은 경우), OLS 추정에 문제가 발생하거나 추정 결과에 대한 신뢰가 떨어지게 된다(5.2절). 그 결과 변수 선택에도 영향을 미치게 된다.
- 예를 들어, 중요한 예측 변수가, 기 포함된 다른 변수의 영향으로 인해, 모형에서 제외되는 경우도 발생하게 된다. 벌점회귀는 모형에서 특정 변수의 과다한 기여를 제한함으로써(큰 모수 값에 벌점을 부여함으로써), 상관된 다른 변수가 동시에 모형에 포함될 수 있도록 허용하는 회귀 방법이다.

8.1 서론

- 벌점회귀 추정량들은 불편성을 만족하지는 않으나, OLS 추정에 비해 MSE를 작게 할 수 있으며, 능형회귀를 제외한 나머지 방법(LASSO, Elestic Net)은 추정 과정에서 변수선택의 기능을 함께 가진다.

8.2 능형회귀

- 능형회귀(ridge regression)는 회귀계수를 다음의 식

$$\hat{\beta} = (X^TX + kI)^{-1}X^Ty$$

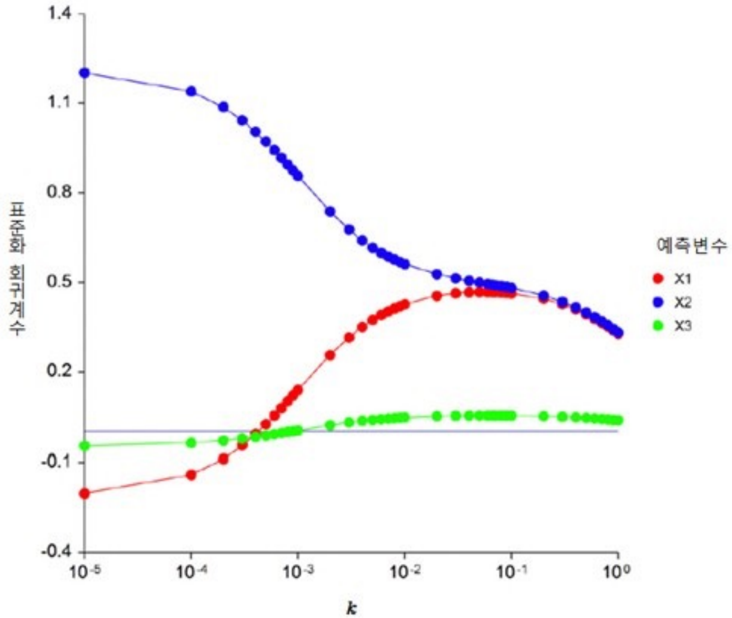
을 통해 추정한다. 여기서 k는 능형모수(ridge parameter)로 불리며 0과 1사이의 값을 가지며 (보통 0.3 보다 작은 값을 가짐), I는 단위행렬을 의미한다. 추정된 능형회귀계수는 편향추정량 (biased estimator)이지만, OLS 추정량 보다 MSE(=분산+제곱편향)를 더 작게 해주는 값이 존재함이 밝혀져 있다.

8.2 능형회귀

- 능형회귀를 수행할 때 한 가지 주의할 점은 모든 예측변수와 반응변수는 표준화(standardized)가 먼저 수행되어야 한다. 위 식에서 와 역시 표준화된 것으로 약속한다. 능형회귀는 상관행렬의 대각원소에 작은 값을 더해줌으로써 수행되는데, 그 명칭은 1의 값을 가지는 상관행렬의 대각원소가 ridge(융기)로 생각될 수 있음에 유래한다.
- 능형회귀의 추정에서 중요한 문제는 값의 결정이다. 일반적으로 능형추적(Ridge Trace), GCV(generalized cross validation)와 Mallow's C_p가 사용된다. 이 가운데 Ridge Trace는 능형모수(λ) 값의 변화에 따른 추정된 모수 값(능형 회귀계수)의 변화를 그림(ridge trace curve)으로 제공한다.
- k값이 증가하면 모든 회귀계수는 안정화 단계를 거쳐 궁극적으로 0으로 수렴하게 되는데, 회귀계수들이 안정화되는 가능한 최소의 k값(이때 회귀계수의 편향이 가장 작아짐)을 선택한다.

8.2 능형회귀

- 예를 들어, 다음의 [그림 8.1]의 능형추적(ridge trace) 곡선에서 회귀계수가 안정화되기 시작하는 k값으로 0.01 보다 조금 큰 값(0.07 정도)이 선택된다. 아래 그림에서는 나타나지 않았지만 각 선의 맨 좌측값($k = 0$인 경우)은 OLS 추정값을 나타낸다.



[그림 8.1] 능형추적 곡선을 이용한 능형모수의 선택

8.2 능형회귀

- 다음의 [예제 1]은 R 패키지 {MASS} 의 lm.ridge() 함수는 능형회귀를 수행한다.

예제 1

longley{datasets} 자료를 이용하여 능형회귀를 수행한다. longley 자료는 다중공선성 문제에 자주 등장하는 거시경제 관련 자료로 7개의 경제변수를 16년(n=16)에 걸쳐 측정된 자료이다.

```
> library(MASS)
> data(longley)
> head(longley)
        y     GNP Unemployed Armed.Forces Population Year Employed
1947 83.0 234.289      235.6        159.0    107.608 1947   60.323
1948 88.5 259.426      232.5        145.6    108.632 1948   61.122
1949 88.2 258.054      368.2        161.6    109.773 1949   60.171
```

8.2 능형회귀

```
> tail(longley,2)
         y     GNP Unemployed Armed.Forces Population Year Employed
1961 115.7 518.173      480.6        257.2    127.852 1961   69.331
1962 116.9 554.894      400.7        282.7    130.081 1962   70.551
```

- 먼저 OLS 추정을 실시한다.

```
> names(longley)[1]<-"y"
> lm.ridge(y~.,longley); #OLS
                     GNP    Unemployed  Armed.Forces    Population
2946.85636017 0.26352725    0.03648291    0.01116105   -1.73702984
         Year   Employed
  -1.41879853 0.23128785
```

8.2 능형회귀

- lm.ridge{MASS} 함수를 이용하여 능형회귀를 수행한다.

```
> r<-lm.ridge(y~.,data=longley,lambda=seq(0,0.1,0.001),model=TRUE)
> r$lambda[which.min(r$GCV)]
[1] 0.006
> r$coef[,which.min(r$GCV)]
        GNP  Unemployed Armed.Forces  Population        Year    Employed
 16.9874524   1.7527228    0.4423901  -8.9474628   1.1782609  -0.1976319
```

- 위의 결과에서 GCV 기준에서 적절한 능형모수로 0.006이 선택되었다.

8.2 능형회귀

예제 2

[예제 1]의 자료에 대해 linearRidge{ridge} 함수를 이용하여 능형회귀를 수행한다.

lambda="automatic"은 Cule(2012)에 의한 능형모수의 선택 방법을 사용한 것으로 0.01이 선택되었다. 이 값은 위의 lm.ridge()를 수행한 결과와 크게 다르지 않다.

```
> names(longley)[1]<-"y"
> mod <-linearRidge(y~.-1, data=longley, lambda="automatic")
> summary(mod)
Call:
linearRidge(formula = y ~ . - 1, data = longley, lambda = "automatic")

Coefficients:
             Estimate Scaled estimate Std. Error (scaled) t value
(scaled) GNP  0.04338        16.69895             3.68931   4.526
Unemployed    0.01184         4.28639             2.50693   1.710
Armed.Forces  0.01381         3.72087             1.90482   1.953
Population   -0.02831        -0.76273             5.28549   0.144
Year          0.65665        12.10811             2.69073   4.500
Employed      0.67454         9.17494             4.99599   1.836
```

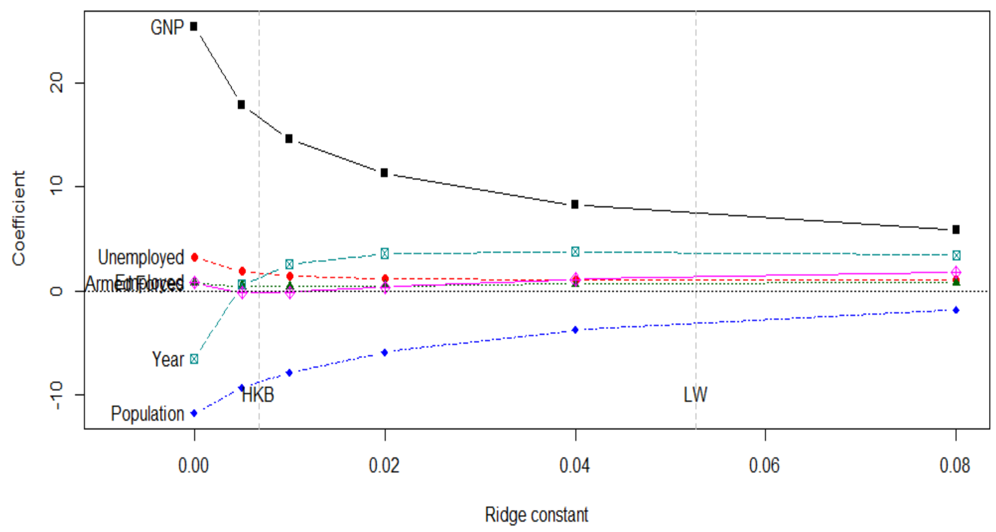
8.2 능형회귀

```
Pr(>|t|)
GNP           6.0e-06 ***
Unemployed     0.0873 .
Armed.Forces   0.0508 .
Population     0.8853
Year          6.8e-06 ***
Employed       0.0663 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Ridge parameter: 0.01046912, chosen automatically, computed using 2
PCs

Degrees of freedom: model 3.67 , variance 3.218 , residual 4.123
```

8.2 능형회귀

- R 패키지 {genridge}의 traceplot() 함수를 이용하여 ridge trace 곡선을 그릴 수 있다.

```
> library(genridge)
> lambda <- c(0, 0.005, 0.01, 0.02, 0.04, 0.08)
> r <- ridge(y~.,longley, lambda=lambda)
> traceplot(r)
```



8.3 벌점회귀 – 능형, LASSO, Elastic Net 회귀

- 다중회귀에서 최소제곱법에 의한 추정치는 주어진 자료에 대해서는 최적이라 할 수 있다. 다만 추정된 회귀식이 새로운 자료에 대해서도 좋은 성능을 유지하기 위해서는 일반화에 유리한 모형이 필요하다.
- 벌점회귀(penalized regression)는, 모형의 일반화(generalization)의 성능이 우수한 회귀 방법으로, 모수에 대해 적절한 제약(regularization) 조건하에서 최소제곱법을 이용하는 방법을 사용한다. 제약조건의 예로는 (i) 회귀계수들의 제곱합이 특정 값 이하이도록(능형회귀), (ii) 회귀계수들의 절대 값의 합이 특정 값 이하이도록(LASSO 회귀), (iii) (i)과 (ii)의 혼합된 형태의 제약(Elastic net)을 주는 방법이 있다.

8.3 벌점회귀 – 능형, LASSO, Elastic Net 회귀

- 다양한 제약조건하에서 오차제곱합을 최소화하는 회귀계수를 찾는 문제는 라그랑지승수법 (Lagrange multiplier method)을 통해 해결된다. 다중회귀의 경우 라그랑지승수법의 목적함수는

$$Q(\beta,\lambda) = \sum_{i=1}^{n}(y - \beta_0 - \sum_{j=1}^{p} x_j\beta_j)^2 + \lambda P(\beta) \quad (\lambda \geq 0)$$

이고, 우리는 이를 최소화하는 회귀계수를 찾는 것이 목적이다.

- 위 식의 두 번째 항은 벌점항(penalty term)으로 불리며, 벌점의 크기를 조절하는 모수 λ(이를 조율모수(tuning parameter)라고 함)와 벌점함수(penalty function) $P(\beta)$로 주어진다.

8.3 벌점회귀 – 능형, LASSO, Elastic Net 회귀

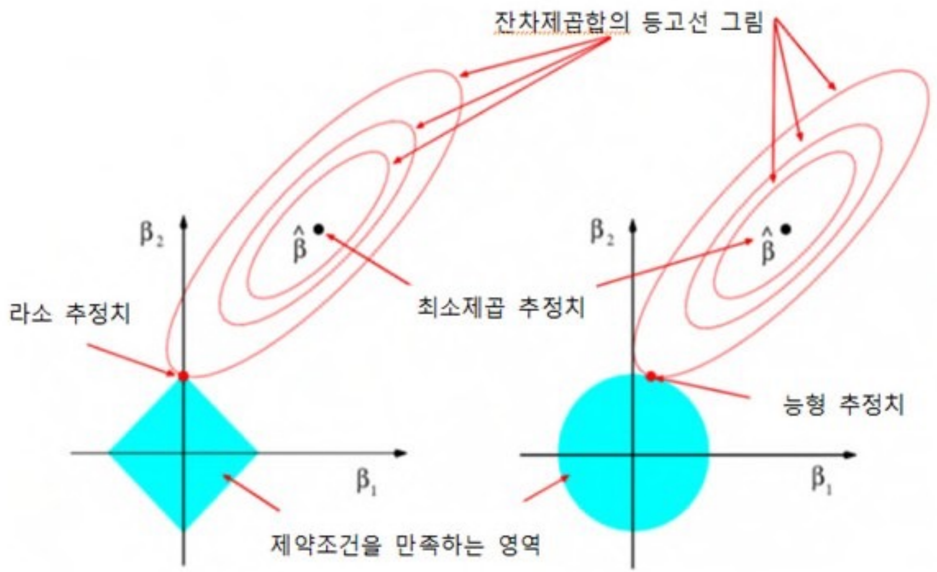
- 벌점함수 $P(\beta)$ 는 다음과 같이 주어진다.
  - **능형회귀의 경우**: $P(\beta) = \sum_{j=1}^{P} \beta_j^2 \quad (l_2 - norm)$
  - **LASSO 회귀의 경우**: $P(\beta) = \sum_{j=1}^{P} |\beta| \quad (l_1 - norm)$

    (LASSO: Least Absolute Shrinkage and Selection Operator)
  - **Elastic net 경우**: $P(\beta) = \sum_{j=1}^{p} \{\alpha|\beta| + (1-\alpha)\beta^2\} \quad (0 \le \alpha \le 1)$

    ($l_1 - norm$과 $l_1 - norm$의 가중합)

8.3 벌점회귀 – 능형, LASSO, Elastic Net 회귀

- 다음의 [그림 8.2]는 LASSO와 능형회귀에서, 제약조건에 따른, 모수추정과 그 차이를 보여준다. 즉, LASSO의 경우(좌측 그림)에는 특정 모수의 추정값이 영(0)이 될 수 있어 자동적으로 변수 선택이 이루어지는 반면, 능형회귀의 경우(우측 그림)에는 변수선택의 기능이 없음을 알 수 있다.



[그림 8.2] LASSO와 능형회귀에서의 모수추정

8.3 벌점회귀 – 능형, LASSO, Elastic Net 회귀

참 고

(1) 위의 목적함수에서 흔히 라그랑지승수(Laglannge mutiplier)로 불리는 λ는 제약모수 (regularization parameter), 벌점함수 $P(\beta)$는 제약자(regularizer)로도 불린다.

(2) Elastic Net 에서 $\alpha = 0$ 이면 능형회귀, $\alpha = 1$ 이면 LASSO 회귀가 된다.

- 벌점회귀는 위의 목적함수 Q를 최소화하는 모수를 찾는 것으로 λ가 영(zero)일 때는 통상적인 최소제곱법의 결과와 동일하나, λ가 커질수록 목적함수를 최소화하는 회귀계수는 더 작은 값으로 축소(shrinkage)되어 추정 된다(실제 자료 분석에서 적절한 값의 선택은 교차타당성 방법 등을 이용한다).

8.3 벌점회귀 – 능형, LASSO, Elastic Net 회귀

- 최소제곱법에서는 특정 변수의 계수가 크게 추정되면(영향력이 커지면), 이와 상관이 높은 다른 변수의 영향은 줄어들게 되어 모형에서 제거되는 경우가 발생한다. 벌점회귀에서는 특정회귀계수가 과다하게 크게 추정되지 않도록 하여(벌점항을 통한 축소 추정), 모형에서 특정 변수의 영향력을 줄여줌으로써 보다 많은 변수들이 모형에 포함될 수 있도록 한다.
- 또한 벌점회귀는 보다 안정적인 추정을 제공하여, 추정된 회귀식이 새로운 자료에 대해 보다 나은 일반화가 되는데 기여한다.

8.3 벌점회귀 – 능형, LASSO, Elastic Net 회귀

- 벌점회귀의 특징을 정리하면 다음과 같다.
  - 다중회귀에서 전통적인 변수선택법(전진선택법, 후진제거법 등)은 변수들을 불연속적으로 선택하지만 (선택 또는 미선택), 벌점회귀에서는 축소추정의 과정을 통해 보다 유연하게 변수를 참여시킨다.
  - 능형회귀는 회귀모수에 대해 축소추정(shrinkage estimates)과 함께 모든 예측변수를 모형에 참여시 키며 따라서 변수선택의 기능은 없다. 반면, LASSO 회귀는 축소추정과 변수선택의 기능을 동시에 가진 다. 즉, 중요하지 않은 변수의 회귀계수가 0으로 추정될 수 있는 여지를 두어 자동적으로 변수선택이 이루 어지게 한다.
  - Elastic Net은 능형회귀와 LASSO 회귀의 절충적인 방법이다. LASSO는 상관이 높은 예측변수가운데 일부만을 선택하는 경향을 가지나, Elastic Net은 상관이 높은 예측변수들이 모두 선택되도록 하는 특징(장점)을 가진다. 이를 그룹화효과(grouping effect)라고 한다.
  - 최소제곱추정량이 최량선형불편추정량(BLUE)을 만족하는 반면, 벌점회귀의 추정량들은 최소제곱추정 량의 축소추정량(shrinkage estimator)으로 편의(bias)는 존재하나 분산(variance)을 작게해 줌으로써 평균제곱오차(MSE)의 측면에서 최소제곱추정량을 개선할 수 있다.

8.3 벌점회귀 – 능형, LASSO, Elastic Net 회귀

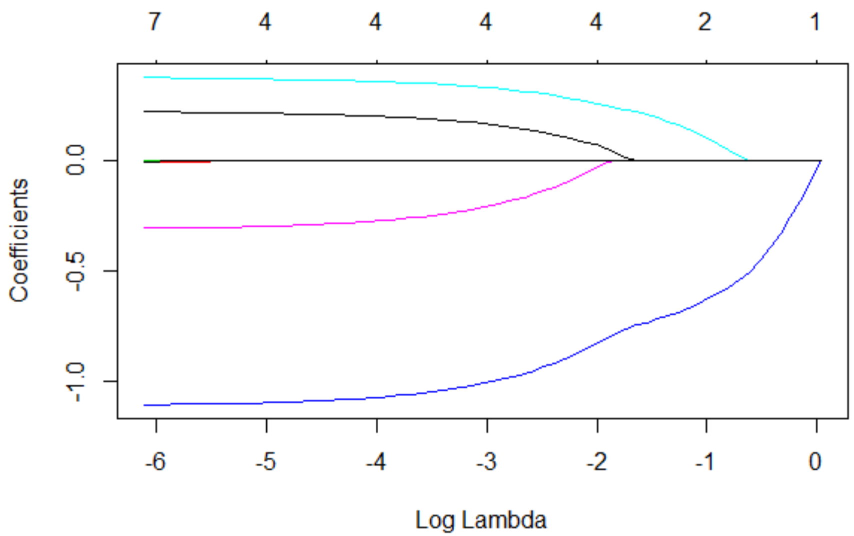
예제 3 R 패키지 {glmnet}을 이용하여 벌점회귀를 수행한다. 분석에는 state.x77 자료를 사용한다.

```
> library(glmnet)
> data(state)
> state <- data.frame(state.x77, row.names=NULL)
> head(state)
  Population Income Illiteracy Life.Exp Murder HS.Grad Frost   Area
1       3615   3624        2.1    69.05   15.1    41.3    20  50708
2        365   6315        1.5    69.31   11.3    66.7   152 566432
3       2212   4530        1.8    70.55    7.8    58.1    15 113417
4       2110   3378        1.9    70.66   10.1    39.9    65  51945
5      21198   5114        1.1    71.71   10.3    62.6    20 156361
6       2541   4884        0.7    72.06    6.8    63.9   166 103766

> y <- state$Life.Exp
> x <- scale(state[,-4])       # 예측변수를 표준화 함
```

8.3 벌점회귀 – 능형, LASSO, Elastic Net 회귀

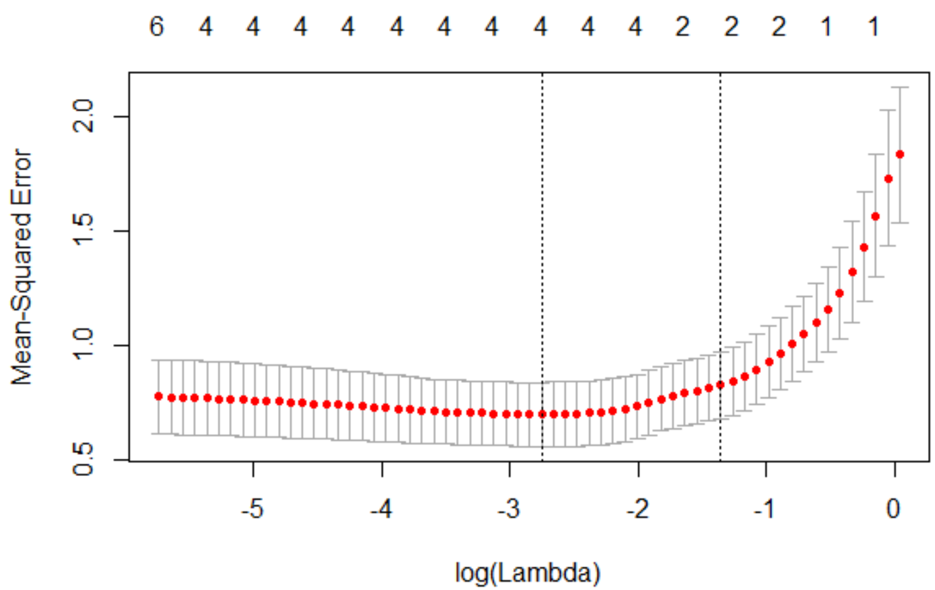
- glmnt() 함수를 이용하여 LASSO 회귀를 수행한다. alpha=1은 LASSO, alpha=0은 능형회귀이다.

```
> lasso <- glmnet(x, y, alpha=1)
> ls(lasso)
> plot(lasso, xvar="lambda")
```



8.3 벌점회귀 – 능형, LASSO, Elastic Net 회귀

```
> cv.lasso <- cv.glmnet(x, y, alpha=1)
> plot(cv.lasso)
```



8.3 벌점회귀 – 능형, LASSO, Elastic Net 회귀

```
> lambda <- cv.lasso$lambda.min
> lambda
[1] 0.09237471
```

해 석

lambda.min은 평균 교차타당오차를 최소로하는 λ이다. cv.glmnet() 함수를 실행할 때마다 다른 값을 제공하므로 유의할 것(이후의 추정 결과도 달라짐에 유의) . cv.glmnet() 함수를 실행하기 전 set.seed() 함수를 사용하면, 일정한 결과를 얻을 수 있다.

8.3 벌점회귀 – 능형, LASSO, Elastic Net 회귀

```
> # 추정된 회귀 계수
> coef <- coef(lasso, s=lambda)
> coef
8 x 1 sparse Matrix of class "dgCMatrix"
                     1
(Intercept) 70.8786000
Population   0.1200443
Income       .
Illiteracy   .
Murder      -0.9141759
HS.Grad      0.2967247
Frost       -0.1164829
Area         .
```

8.3 벌점회귀 – 능형, LASSO, Elastic Net 회귀

해 석

LASSO 회귀를 수행한 결과 4 개의 변수가 선택되었으며, 선택된 변수는 다중회귀의 결과와 동일함을 확인할 수 있다. 여러 가지 벌점회귀를 수행한 결과를 요약하면 다음의 [표 8.1]과 같다. 벌점회귀를 통한 회귀계수의 추정치들이 다중회귀의 결과보다 축소되어 추정되었음을 알 수 있다.

8.3 벌점회귀 – 능형, LASSO, Elastic Net 회귀

[표 8.1] 여러 가지 벌점회귀를 통한 모수추정 결과

예측변수	LASSO (alpha=1)	Ridge (alpha=0)	Elastic Net (alpha=0.5)	Stepwise
(Intercept)	70.8786	70.8786	70.8786	70.8786
Population	0.1200	0.1658	0.1405	0.2238
Income	.	0.0252	0.0405	.
Illiteracy	.	-0.0723	-0.1021	.
Murder	-0.914	-0.9185	-0.8366	-1.1080
HS.Grad	0.296	0.3621	0.3476	0.3762
Frost	-0.116	-0.2511	-0.2232	-0.3089
Area	.	-0.0437	-0.0577	.
비고	예측변수를 표준화한 표준화 회귀계수임			

8.3 벌점회귀 – 능형, LASSO, Elastic Net 회귀

- 새로운 자료에 대한 예측값은 predict() 함수를 이용한다.

```
> predict(cv.lasso, newx= , s="lambda.min")
```

8.3 벌점회귀 – 능형, LASSO, Elastic Net 회귀

참 고 **라소의 일반화(generalization): 그룹라소**

라소의 변형에는 여러 가지가 있다. 이들은 모두 공변량들 간의 서로 다른 유형의 의존성을을 고려하는 방법에 중점을 두고 있다. Elastic net은 예측자의 수가 표본의 크기보다 클 때 능형회귀-형 벌점을 추가하여 성능을 향상시키는 방법으로, 강하게 상관된 변수를 함께 선택하여 전반 적인 예측 정확도를 향상시키는 방법이다.

그룹라소(group lasso)는 관련된(미리 정해진) 공변량의 그룹이 하나의 단위로 선택되도록 허용하는데, 몇 개 공변량이 다른 공변량 없이 포함되는 것이 무의미한 경우에 유용한 방법이다. 예를 들어, 범주형 변수의 수준이 여러 개의 이진 공변량으로 코딩된 경우 또는 생물학에서 어떤 그룹핑이 더 자연적인지를 알고자 하는 경우(유전자와 단백질은 종종 알려진 경로(pathways) 에 놓여 있기 때문에 연구자는 특정 개별 유전자 보다는 어떤 경로가 출현과 관련이 있는지에 더 관심이 있음)가 있다.

8.3 벌점회귀 – 능형, LASSO, Elastic Net 회귀

참 고 **라소의 일반화(generalization): 그룹라소**

그룹 라소의 확장으로, (각 그룹 부공간에 추가적으로 벌점을 추가하는 방법을 사용하여) 개별 그룹 내에서 변수 선택을 수행하는 방법(희박 그룹라소: sparse group lasso)과 공변량의 그룹 간 중복을 허용하는 방법(중복 그룹라소: overlap group lasso)이 있다.

융합라소(fused lasso)는 문제의 공간적 또는 시간적 특성을 설명할 수 있기 때문에 연구되는 시스템의 구조와 더 잘 맞는 추정이 가능하다.

일반화된(regularized) 라소모형은 부경사법, LARS(least-angle regression)와 근접경사법 등을 사용하여 적합 할 수 있다. 제약모수에 대한 최적값은, 모형이 잘 작동하는지 확인하는 데 중요한 부분으로, 일반적으로 교차타당성 방법을 통해 선택된다.