YBIGTA Data Science Team

ISL Study

Section 6: Linear Model Selection and Regularization

2018/02/03

Contents

6. Linear Model Selection and Regularization

반응변수 Y와 설명변수 $X_1, X_2, \cdots, X_p$사이의 상관관계를 설명하기위해, 우리는 다음과 같이 가장 쉽고 간단하다고 할 수 있는 선형회귀모형을 사용했었다.

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \tag{1}$$
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$$

3장에서 학습했듯이 이러한 모형은 보통 최소제곱법을 사용하여 적합한다.

뒤에 이어질 7장과 8장에서는 식 (1)을 더욱 일반화한 비선형적인 모형들을 배우게 된다.

그런데 식 (1)과 같은 선형모형은 비선형방법들과 비교하더라도 추론(inference)의 관점에서 분명한 장점이 있고, 현실적인 문제에서도 경쟁력이 있기 때문에 비선형적 내용으로 넘어가기 전에 일반적인 **최소제곱적합을 다른 적합절차로 대체**하여 단순선형모형을 개선할 수 있는 몇 가지 방법을 논의한다.

Q: 최소제곱적합 대신에 다른 적합절차를 사용하려는 이유는 무엇일까?

A: 우리가 앞에서 배운 Ordinary Least Squares Regression은 모형적합 직전에 설정해주어야 하는 매개변수인 Hyperparameter가 존재하지 않는 모형이다. 이 말인 즉슨 Training Data Set으로 최소제곱적합시킨 우리의 모형은, 적합된 모형 그대로 Test Data Set 예측에 사용해야한다. 다시말해 추가적인 skill을 사용하지 않는 이상 모형개선의 여지는 없는셈.

$$\underbrace{E\left(Y - \hat{f}(X)\right)^2}_{\text{Test MSE}} = \underbrace{\text{Var}\left(\hat{f}(X)\right)}_{\text{Variance}} + \underbrace{\cancelto{0}{\left[E\left(\hat{f}(X)\right) - f(X)\right]^2}}_{\text{Bias}^2} + \underbrace{\text{Var}(\epsilon)}_{\text{irreducible error}}$$

이런 단점을 보완하기 위해 적합단계에서 최소제곱적합이 아닌 다른 적합절차를 사용하거나, 적합이후 변수선택과 같은 추가적인 skill을 사용한다. 그러면 더 나은 **예측 정확도(prediction accuracy)**와 **모델 해석력(model interpretability)**을 얻을 수 있다. 예측 정확도와 모델 해석력이란 다음을 의미한다.

- **예측 정확도(prediction accuracy)**: 반응변수와 설명변수들 사이의 실제 함수관계 ($Y = f(X) + \epsilon$에서 $f(X)$)가 거의 선형적인 경우 최소제곱 추정치들은 편향(bias)이 작을 것이다 (실제 함수관계가 완벽한 선형이면 편향(bias)은 0.)
  만약 $n \gg p$이면, 즉 관측치의 수 n이 변수의 수 p보다 훨씬 크면 최소제곱 추정치들은 낮은 분산(variance)을 가지는 경향이 있고, 따라서 test data set에 대해서도 좋은 성능을 낼 것이다.
  하지만 n이 p보다 아주 크지는 않으면 최소제곱적합에 많은 변동이 존재할 수 있어

과적합을 초래하고, 이로 인해 test data set에 대한 예측결과가 좋지 않을 것이다.
만약 $p > n$이면 더이상 유일한(unique) 최소제곱 계수 추정치가 존재하지 않는다. 즉, 분산이 무한대가 되어 최소제곱 방법을 사용할수 없게 된다.
이 때 추정된 계수들을 **제한(constraining)** 또는 **수축(shrinking)**시키면 편향(bias)은 이전보다 조금 증가하겠지만 분산(variance)을 상당히 감소시킬 수 있다[1]. 따라서 test data set에서의 예측 정확도를 상당히 개선할 수 있다.

- **모델 해석력(model interpretability)**: 다중회귀모형에서 사용되는 일부 또는 많은 변수들은 사실상 반응변수와 연관성이 크지 않은 경우가 많다. 이렇게 관련이 없는 변수들을 포함하는 것은 모형을 불필요하게 복잡하게 만든다. 이런 변수들을 애초에 제외하여 모형을 적합하거나, 불필요한 변수에 대응하는 계수 추정치를 0으로 만들어주는 적합절차를 사용하면 반응변수 Y와 설명변수들의 관계를 좀더 쉽게 해석하는 모형을 얻을 수 있다.
  최소제곱법으로는 정확하게 0인 계수 추정치를 얻게될 가능성은 거의 없다. 따라서 6장에서는 다중회귀모형으로부터 관련없는 변수들을 제외하는, 즉 자동으로 변수 선택(feature selection; variable selection)을 수행하는 몇 가지 기법들을 살펴본다.

식 (1)을 적합하는 데 최소제곱법 대신 사용할 3가지의 대안을 살펴보자.

1. **부분집합 선택(Subset Selection)**: 이 방법은 p개의 설명변수 중에서 반응변수와 관련이 있다고 생각되는 subset을 식별하는 것이다. 그 다음에 변수의 수가 줄어든 subset에 최소제곱법을 사용하여 모형을 적합한다.
   예) Best Subset Selection, Forward Stepwise Selection, Backward Stepwise Selection, Hybrid Approaches

2. **수축(Shrinkage) 또는 정규화(Regularization)**: 이 방법은 p개의 설명변수 모두를 포함하는 모델을 적합시키지만, 일반적인 최소제곱법과는 다르게 계수 추정치가 '0에 가깝게(→Ridge regression) 또는 0으로 정확히(→Lasso regression) ' 수축(Shrinkage) 된다. 이런 수축(Shrinkage)은 모형의 분산을 줄이는 효과가 있고, 변수선택의 한 방법이기도 하다.
   예) Ridge Regression, Lasso Regression

3. **차원축소(Dimension Reduction)**: 이 방법은 p개의 설명변수를 M차원 부분공간으로 projection하는 것이다. (여기서 $M < p$이다). 그 다음 M개의 projection된 설명변수들은 최소제곱법으로 선형회귀모형을 적합하는데 사용.

위 3가지 방법들은 선형회귀모형 뿐만 아니라 4장의 Classification 모형에도 적용 가능하다.

[1] 제한(constraining) 또는 수축(shrinking)은 모형의 유연성(flexibility)을 떨어뜨리므로 모형의 편향(bias)은 증가하고 분산(variance)은 감소하게된다.

6.1 부분집합 선택 (Subset Selection)

6.1.1 Best Subset Selection

반응변수 Y와 p개 설명변수들의 관계를 가장 잘 설명해주는 설명변수들의 Best Subset을 찾기 위해, p개 설명변수의 모든 가능한 조합 각각에 대해 최소제곱회귀를 적합한다. 즉, 설명변수를 하나도 포함하지 않는 모형부터

$$Y = \beta_0 + \epsilon$$

p개 모두 사용하는 모형까지

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

$$\begin{aligned}\sum_{k=0}^{p} \binom{p}{k} &= \binom{p}{0} + \binom{p}{1} + \cdots + \binom{p}{p} \\ &= 2^p\end{aligned}$$

총 2^p개의 가능한 모형을 전부 검토하여 '최고의 모형'을 찾는 방법이다.

이 과정은 다음의 [Algorithm 6.1]에 나오듯이 2단계로 나누어진다.

[Algorithm 6.1] Best subset selection

1. $\mathcal{M}_0$는 설명변수를 하나도 포함하지 않는 null model이라고 하자. 이 모형은 단순히 반응변수의 표본평균만을 예측해주는 모형.
2. **(Step 1)** $k = 1, 2, \cdots, p$에 대하여,
   (a) 정확히 k개의 설명변수를 포함하는 모든 $\binom{p}{k}$개의 모형을 적합한다.
   (b) $\binom{p}{k}$개의 모형 중 '최고'의 모형을 골라 $\mathcal{M}_k$라 한다. 여기서 '최고'는 가장 작은 RSS값이나 가장 큰 R^2값을 갖는 것으로 정의된다.
3. **(Step 2)** 교차검증(cross validation)된 예측오차(prediction error), C_p (AIC), BIC 또는 adjusted R^2을 사용하여 $\mathcal{M}_0, \cdots, \mathcal{M}_p$ 중에서 '최고'의 모형을 하나 선택한다.

[Algorithm 6.1]의 핵심은 **설명변수의 갯수가 동일한 모형들 중 최고의 모형**(training data에 대해)을 먼저 식별하여, 원래 2^p개의 가능한 모형 중에서 하나를 선택해야 하는 문제를 $p+1$개의 모형 중에서 하나를 선택하는 문제로 축소한 것이다. 이 때, 설명 변수의 갯수가 동일한 (유연성이 같은) 모형들의 비교 기준으로는 RSS값이나 R^2값을 이용한다.

그리고 이렇게 선정된 $p+1$**개 모형들(설명변수의 갯수가 다른 모형들)**의 비교는 **Step 2**에 제시된 다른 기준들을 이용한다. 왜냐하면 모형에 포함된 설명변수의 수가 증가함에 따라 RSS는 단조감소하고, R^2는 단조증가하기 하여 모형비교에 적절한 기준이 아니기 때문이다.

아래 FIGURE 6.1은 $p+1$개 모형을 선정하는 [Algorithm 6.1] 의 Step 1를 보여주고 있다.

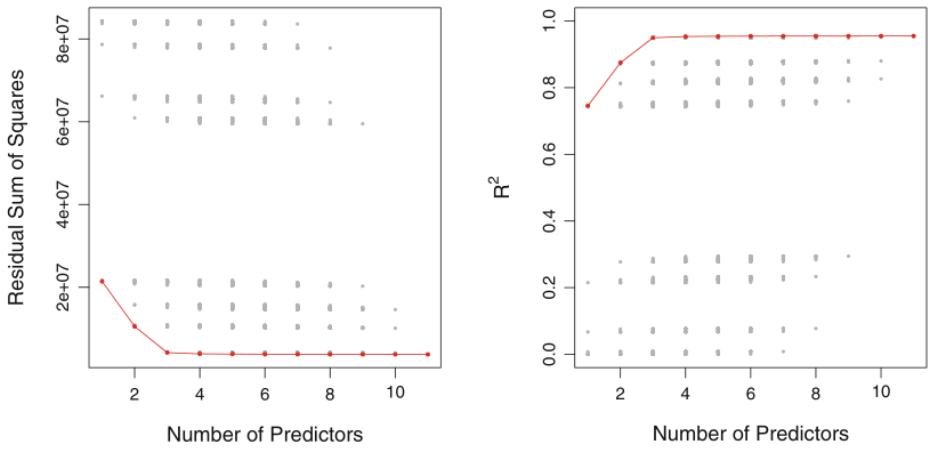


FIGURE 6.1. *For each possible model containing a subset of the ten predictors in the* `Credit` *data set, the RSS and* R^2 *are displayed. The red frontier tracks the* best *model for a given number of predictors, according to RSS and* R^2*. Though the data set contains only ten predictors, the* x*-axis ranges from* 1 *to* 11*, since one of the variables is categorical and takes on three values, leading to the creation of two dummy variables.*

Best Subset Selection은 간단하고 개념적으로도 좋은 기법이지만 계산상의 제약이 있을 수 있다. 왜냐하면, 고려해야 하는 가능한 모형의 수가 p가 증가함에 따라 급격히 늘어나기 때문이다. 즉, $p=10$인 경우 대략 $2^{10}=1,000$개, $p=20$이면 100만개 이상의 모형을 계산해야 한다. 결과적으로 Best Subset Selection은 p가 약 40보다 크면 굉장히 빠른 컴퓨터를 가지고도 계산이 불가능해 진다. 따라서 Best Subset Selection의 대안으로 계산적으로 효율적인 방법들에 대해서 살펴보자.

6.1.2 Stepwise Selection

Forward Stepwise Selection

Best Subset Selection이 2^p개의 가능한 모든 모형들을 고려하는 반면에 Forward Stepwise Selection은 훨씬 적은 수의 모형들을 고려한다. Forward Stepwise Selection은 설명변수가 하나도 포함되지 않은 모형에서 시작하여 모든 설명변수가 모형에 포함될 때까지 한번에 하나씩 설명변수를 추가한다. 특히 각 단계에서 모형에 추가되는 변수는 적합에 가장 큰 추가적 향상을 제공하는 변수이다. Forward Stepwise Selection의 절차는 [Algorithm 6.2] 와 같다.

[Algorithm 6.2] Forward Stepwise Selection

1. $\mathcal{M}_0$는 설명변수를 하나도 포함하지 않는 null model이라고 하자.
2. $k = 0, 1, \cdots, p-1$에 대하여,
   (a) $\mathcal{M}_k$에 하나의 설명변수를 추가한 모든 $p-k$개의 모형을 고려한다.
   (b) $p-k$개의 모델 중에서 '최고'를 골라 $\mathcal{M}_{k+1}$이라 한다. 여기서 '최고'는 가장 작은 RSS나 가장 큰 R^2을 갖는 것으로 정의된다.
3. 교차검증(cross validation)된 예측오차(prediction error), C_p (AIC), BIC 또는 adjusted R^2을 사용하여 $\mathcal{M}_0, \cdots, \mathcal{M}_p$ 중에서 '최고'의 모형을 하나 선택한다.

2^p개의 모형을 적합하는 Best Subset Selection과는 달리, Forward Stepwise Selection은

$$1 + \sum_{k=0}^{p-1}(p-k) = 1 + \frac{p(p+1)}{2}$$

개의 모형을 적합한다. 예를 들어 $p = 20$인 경우 Best Subset Selection은 1,048,576개의 모형을 적합해야 하지만, Forward Stepwise Selection은 211개의 모형만 적합하면 된다.

계산상의 이점을 얻은 대신 잃은것은 무엇일까? Forward Stepwise Selection은 p개 설명변수로 구성할 수 있는 모든 2^p개의 subset모형 중에서 최고의 모델을 찾는다는 보장이 없다. 예를 들어, $p = 3$개의 설명변수를 갖는 데이터셋에서 가능한 최고의 1변수 모형은 X_1을 포함하고, 최고의 2변수 모형은 X_2와 X_3을 포함한다고 해보자. 그러면 Forward Stepwise Selection은 최고의 2변수 모형을 선택하는 데 실패할 것이다. 왜냐하면, $\mathcal{M}_1$은 X_1을 포함할 것이고, 따라서 $\mathcal{M}_2$도 하나의 추가 변수와 함께 X_1을 포함해야 하기 때문이다.

아래 표는 Credit 자료에서 Best Subset Selection과 Forward Stepwise Selection으로 선택되는 처음 4개의 모형을 보여준다.

변수의 갯수	Best Subset Selection	Forward Stepwise Selection
1	rating	rating
2	rating, income	rating, income
3	rating, income, student	rating, income, student
4	**cards**, income	rating, income
	student, limit	student, limit

Forward Stepwise Selection은 $n < p$인 경우에도 적용할 수 있다. 하지만 이 경우에는 부분모형(submodel)$\mathcal{M}_0, \cdots, \mathcal{M}_{n-1}$만 구성하는 것이 가능하다. 왜냐하면, 각 부분모형은 최소제곱법을 이용하여 적합하므로 $p \geq n$이면 유일한 해가 제공되지 않기 때문이다.

Backward Stepwise Selection

Backward Stepwise Selection 또한 Best Subset Selection에 대한 효율적인 대안을 제공한다. 하지만 Forward Stepwise Selection과 달리 p개의 설명변수 모두를 포함하는 완전모형을 가지고 시작하고, 한번에 하나씩 반복적으로 유용성이 가장 적은 설명변수를 제외한다. 알고리즘은 다음과 같다.

[Algorithm 6.3] Backward Stepwise Selection

1. $\mathcal{M}_p$는 p개의 설명변수 모두를 포함하는 완전 모형(full model)이라 한다.
2. $k = p, p-1, \cdots, 1$에 대하여,
   (a) $k-1$개의 설명변수에 대해, $\mathcal{M}_k$에서 하나의 설명변수를 제외한 모든 k개의 모형을 고려한다.
   (b) k개의 모형 중에서 '최고'를 골라 $\mathcal{M}_{k-1}$이라 한다. 여기서 '최고'는 가장 작은 RSS나 가장 큰 R^2을 갖는 것으로 정의된다.
3. 교차검증(cross validation)된 예측오차(prediction error), C_p (AIC), BIC 또는 adjusted R^2을 사용하여 $\mathcal{M}_0, \cdots, \mathcal{M}_p$ 중에서 '최고'의 모형을 하나 선택한다.

Backward Stepwise Selection도 Forward Stepwise Selection처럼 $1 + \frac{p(p+1)}{2}$개의 모형만 검색하므로 p가 너무 커서 Best Subset Selection을 적용할 수 없는 상황에서도 적용이 가능하다. 그러나 역시 Forward Wtepwise Selection처럼 Backward Stepwise Selection도 p개 설명변수로 구성할 수 있는 모든 2^p개의 subset모형 중에서 최고의 모델을 찾는다는 보장이 없다.

Backward Stepwise Selection은 $n > p$인 경우에만 사용할 수 있다(full model 적합이 가능해야 하므로). 따라서 Best Subset Selection, Forward Stepwise Selection, Backward Stepwise Selection중 $n < p$ 일 때 사용가능한 방법은 Forward Stepwise Selection이 유일하다.

Hybrid Approaches

또 하나의 대안적인 방법으로 Forward Stepwise Selection, Backward Stepwise Selection의 하이브리드 버전이 있다. 이 방법은 변수들이 모형에 순차적으로 추가된다는 점에서 전진단계적 선택과 비슷하지만, 새로운 변수를 추가한 후에 모형 적합을 더 이상 향상시키지 않는 변수가 있으면 제거할 수도 있다. 이러한 접근 방식은 Forward, Backward Stepwise Selection의 계산적 장점은 유지하면서 Best Subset Selection의 선택을 모방하고자 하는 것이다.

6.1.3 최적의 모형 선택(Choosing the Optimal Model)

앞절에서 보았듯이, 설명변수의 수가 다른 부분모형(submodel)들을 비교하여 '최고'의 모형을 선택하기 위해서는 교차검증(cross validation)된 예측오차(prediction error), C_p (AIC), BIC 또는 adjusted R^2을 사용한다고 하였는데 이를 알아보자.

위 방법들은 모두 Test MSE를 추정하는 추정량으로써 추정방식에 따라 다음과 같이 분류 가능하다.

1. 과적합으로 인한 편향(bias)를 고려하도록 Training MSE를 '조정'하여 Test MSE를 **간접적으로 추정**한다.
   ex) C_p (AIC), BIC, adjusted R^2
2. 5장에서 다루었던 validation set approach나 cross-validation approach를 사용해 Test MSE를 **직접 추정**한다.

C_p, AIC, BIC, and Adjusted R^2

d개의 설명변수를 포함하는 적합된 최소제곱 모형에 대해 Test MSE의 C_p 추정값은 다음 식을 사용하여 계산된다.

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

여기서 $\hat{\sigma}^2$은

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \tag{3}$$

위 식 (3)의 각 반응변수 측정값과 연관된 오차 ϵ의 분산에 대한 추정값이다[2]. C_p 통계량은 Test MSE를 과소추정하는 경향이 있는 Training MSE를 조정하기 위해 Training RSS에 $2d\hat{\sigma}^2$의 패널티(penalty)를 더한다. 모형에 포함된 설명변수의 수가 증가할수록 패널티도 명백히 증가하는데, 이것은 훈련 RSS가 감소하는 것을 조정하기 위한 것이다. 결론적으로, C_p 통계량은 낮은 Test MSE를 갖는 모형에 대해 작은 값을 가지는 경향이 있으므로, 모형들의 집합에서 최고의 모형을 결정할 때 가장 낮은 C_p 값을 가지는 모형을 선택한다.

AIC(Akaike information criterion)는 최대가능도 추정법에 의해 적합된 모형들에 대해서만 사용할수 있는 방법이고 다음과 같이 정의된다.

$$\text{AIC} = \widehat{\text{Test MSE}} = -2\log(L) + 2d$$

여기서 L은 가능도함수이고, d는 설명변수의 갯수이다.

[2] 이 책의 범위를 벗어나는 내용이긴 하지만, 만약 $\hat{\sigma}^2$이 σ^2의 비편향추정치이면 C_p는 Test MSE의 비편향추정치임을 보일 수 있다.

특별히 식 (3)과 같은 최소제곱 회귀모형의 경우엔 다음과 같이 AIC를 계산할 수 있으며[3]

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

만약 식 (3)의 오차들이 정규분포를 따르는 모형이라면 최대가능도 추정량과 최소제곱 추정량은 같기 때문에 는 최소제곱 적합시 계산한 RSS와 $\hat{\sigma}^2$를 그대로 사용하여 계산할 수 있다. 이를 통해 오차들이 정규분포를 따르는 회귀모형의 경우는 C_p와 AIC는 서로 비례함을 알 수 있으며 둘 중 하나의 추정량만 선택해서 사용하면 될것이다.

BIC(Bayesian information criterion)는 베이즈 관점에서 파생되었지만 C_p 및 AIC와도 유사하다. 설명변수의 수가 d개인 최소제곱 회귀모형에 대해 BIC는 다음과 같이 주어진다.

$$\text{BIC} = \widehat{\text{Test MSE}} = \frac{1}{n}\left(RSS + \log(n)\cdot d\hat{\sigma}^2\right)$$

C_p와 AIC처럼 낮은 BIC값을 가지는 모형이 선택된다. BIC는 C_p에서 $2d\hat{\sigma}^2$을 $\log(n)d\hat{\sigma}^2$으로 대체한 것이다. 여기서 n은 관측치의 갯수이다. $n > 7$이면 $\log(n) > 2$ 이기 때문에 BIC 통계량은 일반적으로 변수의 수가 많은 모형에 C_p보다 더 심한 패널티를 부여한다. 그 결과 C_p보다 더 작은 크기의 모형이 선택된다. 아래 Figure 6.2의 Credit 자료에 대한 결과를 보면 이것이 사실임을 알 수 있다.

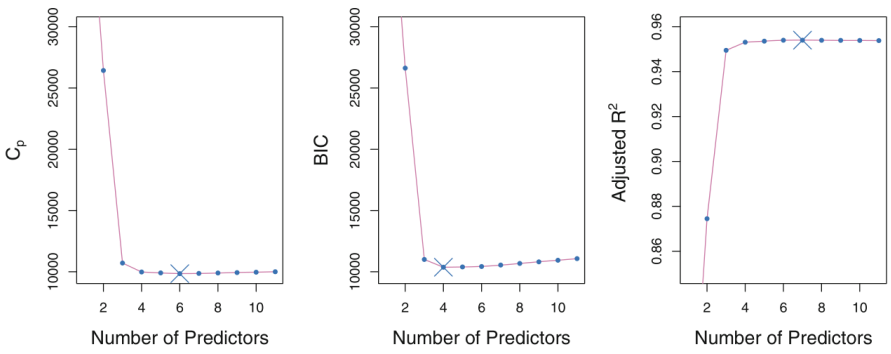


FIGURE 6.2. *C_p, BIC, and adjusted R^2 are shown for the best models of each size for the* `Credit` *data set (the lower frontier in Figure 6.1). C_p and BIC are estimates of test MSE. In the middle plot we see that the BIC estimate of test error shows an increase after four variables are selected. The other two plots are rather flat after four variables are included.*

설명변수의 수가 d개인 최소제곱 회귀모형에 대해, Adjusted R^2 통계량은 다음과 같이 계산된다.

$$\text{Adjusted } R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

[3]정확한 식은 AIC' $= n\ln\left(\frac{\text{RSS}}{n}\right) + 2d$ 이지만, 가장 작은 AIC 를 가지는 모형은 또한 가장 작은 AIC'를 갖기 때문에 표현의 단순함을 위해 위와같이 표현.

작은 값일수록 낮은 Test MSE를 갖는다는 것을 나타내는 C_p, AIC, BIC와 달리, Adjusted R^2는 값이 클수록 모형의 Test MSE가 작다는 것을 의미한다.

Validation and Cross-Validation

5장에서 학습한 검증셋 방법(Validation set approach)과 교차검증 방법(Cross-Validation methods)들은 C_p, AIC, BIC, Adjusted R^2와 비교해 Test MSE를 **직접적으로 추정**하는 통계량이고, 실제 모형에 대한 가정을 적게 한다는 장점이 있다. 이 방법은 또한 더 넓은 범위에 걸쳐 모형을 선택하는 데 사용될 수 있으며, 심지어는 모형의 자유도(예를 들어, 설명변수의 수)를 정확히 알아내기 어렵거나 오차의 분산 σ^2을 추정하기 어려운 경우에도 사용할 수 있다.

예전에는 교차검증을 수행하는 데 p가 크거나 n이 클 경우 계산상 제약이 있어 C_p, AIC, BIC, Adjusted R^2를 주로 사용했었다. 그러나, 요즘은 컴퓨터가 빨라져 교차검증에 필요한 계산은 아무 문제도 아니다. 따라서 교차검증은 고려중인 다수의 모형으로부터 선택을 하기 위한 아주 유용한 기법이다.

Figure 6.3은 Credit 자료에서 최고의 d-변수 모형에 대한 BIC, 검증셋 오차, 교차검증 오차를 d의 함수로써 나타내고 있다.

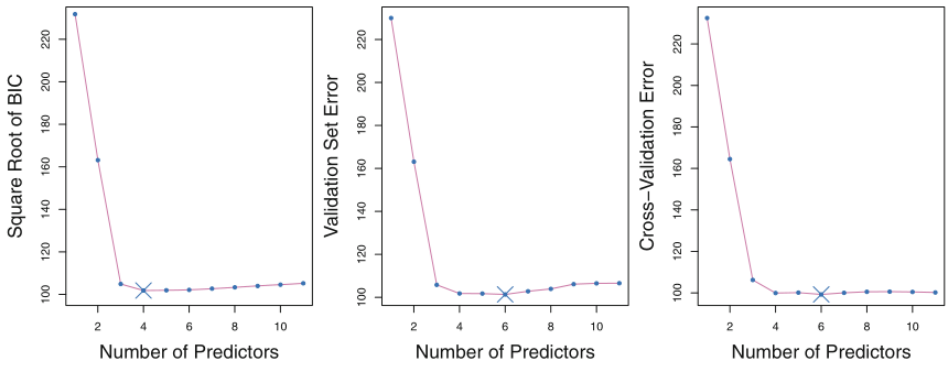


FIGURE 6.3. *For the* `Credit` *data set, three quantities are displayed for the best model containing d predictors, for d ranging from* 1 *to* 11. *The overall* best *model, based on each of these quantities, is shown as a blue cross.* Left: *Square root of BIC.* Center: *Validation set errors.* Right: *Cross-validation errors.*

검증셋 오차는 관측치의 3/4를 training data set으로 선택하고, 나머지를 validation data set으로 선택하여 계산한 것이다. 교차검증 오차는 $k = 10$ fold를 이용해 계산되었다.

이 경우에 검증셋 방법과 교차검증 방법에 의해 선택된 모형은 둘 다 6-변수 모형이다. 하지만, 세 기법 모두 3-변수에서 11-변수 모형들의 추정된 Test MSE는 상당히 유사하다. 또한 training data set과 validation data set 분할을 다르게 하여 검증셋 기법을 반복하거나, fold 크기가 다른 교차검증 방법을 반복한다면 추정된 Test MSE가 가장 낮은 모형은 분명히 바뀔 것이다.

이런 상황에서는 **'one-standard-error' 규칙**을 사용하여 모형을 선택할 수 있다. 먼저 각 모형 크기에 대한 추정된 Test MSE의 표준오차를 계산한다. 그 다음에 추정된 Test MSE 곡선에서 가장 작은 값의 1 표준오차(standard error) 이내에 있는 추정된 Test MSE에 대해 가장 크기가 작은 모형을 선택한다. 그 이유는 모형들이 거의 비슷한 수준이라면 가장 단순한 모형, 즉 설명변수의 수가 가장 적은 모형을 선택하고자 하기 때문이다. 이 예의 경우 검증셋 또는 교차검증 기법에 one-standard-error 규칙을 적용하면 3-변수 모형이 선택된다.

6.2 Shrinkage 방법

6.1절에서 설명한 부분집합 선택(Subset Selection) 방법들은 설명변수들의 subset을 포함하는 선형모형을 적합하는 데 최소제곱법을 사용한다. 이에 대한 대안으로, 계수 추정치들을 **제한(constrains)**하거나 **정규화(regularizes; 규칙화)** 하는 기법을 사용하여 p개의 설명변수 모두를 포함하는 모형을 적합할 수 있다. 이 기법은 계수 추정치들을 0으로 **수축(shrink)**하는 것과 같다. 이러한 제한이 모형의 성능을 왜 향상시키는지 명확하게 바로 드러나지는 않을 수 있지만, 계수 추정치들을 수축하는 것은 추정치들의 분산을 상당히 줄일 수 있는 것으로 밝혀져 있다. 회귀계수들을 0으로 수축시키기 위한 두 가지 가장 잘 알려진 기법은 능형회귀(ridge regression)와 LASSO이다.

6.2.1 능형회귀(ridge regression)

3장에서 학습한 최소제곱 적합 절차는 다음 식을 최소로 하는 값을 이용해 $\beta_0, \beta_1, \cdots, \beta_p$를 추정하였다.

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

능형회귀(ridge regression)는 약간 다른 수량(quantity)을 최소화하여 계수들을 추정한다는 점을 제외하면 최소제곱법과 아주 유사하다. 특히, 능형회귀 계수 추정치 $\hat{\beta}^R$은 다음 식을 최소로 하는 값이다.

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{4}$$

여기서 $\lambda \geq 0$는 별도로 결정되는 **조율 파라미터(tuning parameter; hyperparameter)**이다. 식 (4)는 두 가지 다른 기준을 절충한다. 최소제곱법에서와 같이, 능형회귀(ridge regression)는 RSS를 작게 만들어 데이터에 잘 적합하는 계수 추정치를 찾는다. 하지만 **수축 패널티(shrinkage penalty)**라 불리는 두 번째 항 $\lambda \sum_j \beta_j^2$은 $\beta_1, \cdots, \beta_p$가 0에 가까울수록 작기 때문에, 따라서 β_j의 추정치를 0으로 **수축**하는 효과가 있다. 조율 파라미터 λ는 회귀계수 추정치에 대한 이 두 항의 상대적인 영향을 제어한다. $\lambda = 0$일 때 패널티 항의 영향이 없어

능형회귀(ridge regression)는 최소제곱 추정치를 제공한다. 하지만 $\lambda \to \infty$에 따라 수축 패널티의 영향이 커져 능형회귀 계수 추정치들은 0으로 접근할 것이다.

단 하나의 계수 추정치들의 집합을 생성하는 최소제곱법과 달리, 능형회귀(ridge regression)는 λ의 값 각각에 대해 다른 집합의 계수 추정치 $\hat{\beta}_\lambda^R$를 생성할 것이다. λ에 대한 적절한 값을 선택하는 것은 아주 중요하며, 이에 대해서는 6.2.3절에서 논의한다.

식 (4)를 보면 수축 패널티는 절편 β_0에는 적용되지 않는다. 수축하고자 하는 것은 반응변수에 대한 각 설명변수들의 추정된 연관성이기 때문이다.

Credit 자료에 대한 응용

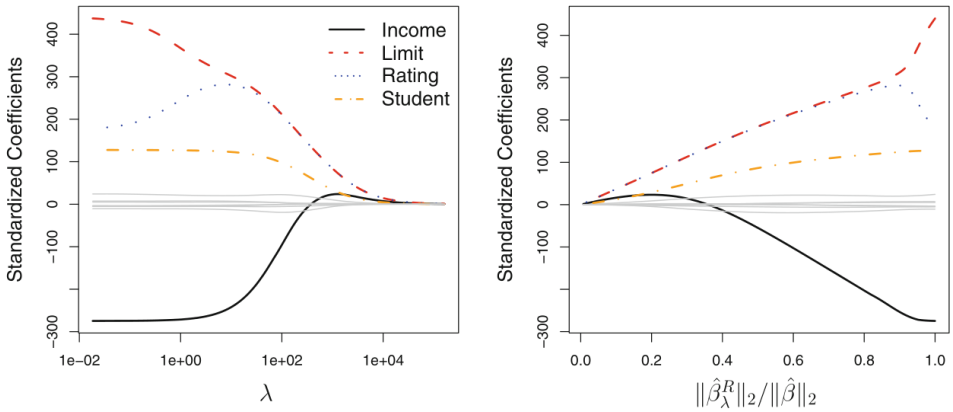


FIGURE 6.4. *The standardized ridge regression coefficients are displayed for the* `Credit` *data set, as a function of* λ *and* $\|\hat{\beta}_\lambda^R\|_2/\|\hat{\beta}\|_2$.

Figure 6.4는 Credit 자료에 대한 능형회귀(ridge regression) 계수 추정치들을 나타낸다. 왼쪽 패널에서 각 곡선은 λ의 함수로 표시되며 10개 변수 중의 하나에 대한 능형회귀 계수 추정치에 해당된다. 그래프의 왼쪽 끝으로 가면 λ가 0이 되고 대응하는 능형 계수 추정치는 최소제곱 추정치와 같아진다. 그러나 λ가 증가함에 따라 능형 계수 추정치들은 0을 향해 수축된다. λ가 아주 클 경우 모든 능형 계수 추정치가 기본적으로 0이 되어 설명변수를 하나도 포함하지 않는 영모형(null model)이 된다.

능형 계수 추정치들은 λ가 증가함에 따라 그 합은 전체적으로 감소하는 경향이 있지만, rating과 income과 같은 개개의 변수들은 λ가 증가하면 간혹 증가할 수도 있다.

Figure 6.4의 오른쪽 패널은 왼쪽 패널과 동일한 능형 계수 추정치들을 나타내지만 x축에 λ대신 $||\hat{\beta}_\lambda^R||_2/||\hat{\beta}||_2$를 사용한다. 여기서 $\hat{\beta}$은 최소제곱 계수 추정치들의 벡터이다. $||\beta||_2$는 벡터의 ℓ_2-norm을 나타내는데, 이는 0에서 β까지의 거리인 $||\beta||_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$으로 정의된다.

λ가 증가함에 따라 $\hat{\beta}_\lambda^R$의 ℓ_2-norm은 **항상** 감소할 것이고 그래서 $||\hat{\beta}_\lambda^R||_2/||\hat{\beta}||_2$도 감소할 것이다. $||\hat{\beta}_\lambda^R||_2/||\hat{\beta}||_2$의 범위는 1에서 0까지이다. 이 값은 $\lambda = 0$일 때(능형회귀 계수 추정치가

최소제곱 추정치와 같아져 그 ℓ_2-norm들이 동일한 경우) 1이고, $\lambda = \infty$일 때(능형회귀 계수 추정치는 ℓ_2-norm이 0인 영벡터인 경우) 0이다. 그러므로, Figure 6.4의 오른쪽 패널에서 x 축은 능형회귀 계수 추정치가 0을 향해 수축된 양으로 생각할 수 있다. 작은 값은 수축된 계수들이 0에 아주 가깝다는 것을 나타낸다.

3장에서 논의된 일반적인 최소제곱 계수 추정치들은 scale equivariant하다. 예를 들어 상수 c를 X_j에 곱하면 최소제곱 추정치는 $1/c$배로 스케일링(scaling)된다. 다르게 말하면, j번째 설명변수가 어떻게 스케일링되는지와 관계없이 $X_j\hat{\beta}_j$은 여전히 동일할 것이다. 이에 반해 능형회귀 계수 추정치는 주어진 설명변수에 어떤 상수를 곱할 때 **현저하게** 바뀔 수 있다. 예를 들어, 달러로 측정되는 income 변수를 고려해보자. 누군가가 1,000달러 단위로 수입을 측정했다고 하면 income 변수의 관측치 값은 1,000배 줄어들 것이다. 능형회귀 공식 (4)에 포함된 계수들의 제곱합 항 때문에, 이러한 스케일 변화는 income에 대한 능형회귀 계수 추정치를 단순히 1,000배 변하게 하는 데 그치지는 않을 것이다. 즉, $X_j\hat{\beta}^R_{j,\lambda}$은 λ값 뿐만 아니라 j번째 설명변수의 스케일링에 따라 다를 것이다.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\hat{\beta}_1^* = \frac{\sum_{i=1}^{n}\left(\frac{x_i}{c} - \frac{\overline{x}}{c}\right)(y_i - \overline{y})}{\sum_{i=1}^{n}\left(\frac{x_i}{c} - \frac{\overline{x}}{c}\right)^2} = \frac{\frac{1}{c}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\frac{1}{c^2}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$= c\hat{\beta}_1 \quad \rightarrow \therefore \text{scale equivariant}$$

	X_1 **(단위 : 1원)** $\longrightarrow$ X_1 **(단위 : 100원)**	
	$\hat{\beta}_1$	$\hat{\beta}_1$
OLS Regression	5	500
ex) 200원이 관측값일 때	$\hat{\beta}_1 X_1 = 5 \times 200 = 1000$	$\hat{\beta}_1 X_1 = 500 \times 2 = 1000$
Ridge Regression	1	??
ex) 200원이 관측값일 때	$\hat{\beta}_1 X_1 = 1 \times 200 = 200$	$\hat{\beta}_1 X_1 = ? \times 2 \neq 200$

"근데... 우리가 가지고 있는 training data set에서 설명변수들의 측정 scale을 바꾸지 않고 계속 그대로 사용 할꺼라면 ridge regression적합을 위해 굳이 표준화 시켜줄 필요가 없는거 아냐?"

ㄴㄴㄴ!!!, $X_j\hat{\beta}^R_{j,\lambda}$의 값은 심지어 다른 설명변수들의 스케일링에 따라 달라지는 경우도 있다!
따라서 능형회귀(ridge regression)는 설명변수들이 모두 동일한 스케일을 가지도록 아래 식을

사용해 표준화한 다음에 적용하는 것이 가장 좋다.

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}} \tag{5}$$

표준화된 설명변수들을 이용하면 최종 적합은 설명변수들이 측정된 스케일에 의존적이지 않을 것이다. Figure 6.4에서 y축은 표준화된 능형회귀 계수 추정치를 나타내는데, 이것은 표준화된 설명변수들을 사용해 능형회귀를 수행한 결과로 얻은 것이다.

능형회귀가 최소제곱보다 나은 이유

최소제곱에 대한 능형회귀의 장점은 bias-variance trade-off에 원인이 있다.

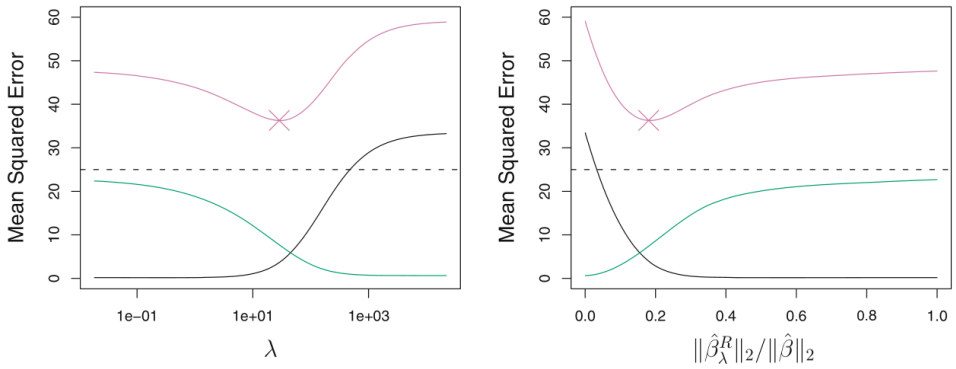


FIGURE 6.5. *Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of* λ *and* $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$. *The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.*

$$\underset{\text{Test MSE}}{E\left(Y - \hat{f}(X)\right)^2} = \underset{\text{Variance}}{\text{Var}\left(\hat{f}(X)\right)} + \underset{\text{Bias}^2}{\left[E\left(\hat{f}(X)\right) - \hat{f}(X)\right]^2} + \underset{\text{irreducible error}}{\text{Var}(\epsilon)}$$

λ가 증가하면 능형회귀 적합의 유연성이 감소하게 되어 분산(variance)은 감소하지만 편향(bias)는 증가한다. 그러나 λ가 증가함에 따라 능형 계수 추정치의 수축은 편향(bias)을 약간 증가시키지만 예측값의 분산(variance)을 현저하게 줄일 수 있다.

Figure 6.5의 왼쪽 패널에서 $\lambda = 10$정도 될 때, 검은색으로 그려진 편향(bias)은 $\lambda = 0$일 때 (최소제곱적합) 보다 아주 조금 증가했지만, 분산(variance)은 급격히 낮아졌음을 확인할 수 있다. 최소 MSE는 대략 $\lambda = 30$일 때 얻어진다. 흥미롭게도, 예측의 높은 분산 때문에, $\lambda = 0$일 때 (최소제곱적합) Test MSE는 $\lambda = \infty$일 때 모든 계수 추정치가 0인 null model의 Test MSE만큼이나 높다. 하지만 적절한 λ값에 대해서는 Test MSE가 상당히 낮다.

6.2.2 Lasso

능형회귀(ridge regression)는 한 가지 분명한 단점이 있다. 앞서 논의했던 변수선택법들과는 달리, 능형회귀는 최종 모형에 p개 설명변수 모두를 포함할 것이다. 식 (4)에서 패널티 $\lambda \sum \beta_j^2$은 모든 계수를 0을 향해 수축시킬 것이지만 계수 중 어떤 것도 ($\lambda = \infty$가 아니라면) 정확하게 0으로 만들지는 않을 것이다. 이것은 예측 정확도에 있어서는 문제가 되지 않을 수도 있지만 변수의 수 p가 상당히 클때는 모형을 해석하는 데 어려움을 초래할 수 있다. 예를 들어, Credit 자료에서 변수선택법에 의해 가장 중요한 변수들은 income, limit, rating, student 이였으므로, 이 변수들만 포함하는 모형을 만들고자 할 수 있다. 하지만 능형회귀는 항상 10개의 설명변수 모두를 포함하는 모형을 생성할 것이다. λ값의 증가가 계수들의 크기를 줄이는 경향이 있겠지만 어떤 변수들을 제외한 결과를 제공하지는 않을 것이다.

Lasso(Least Absolute Shrinkage and Selection Operator)는 능형회귀의 이러한 단점을 극복하는 비교적 최신(?)[4] 기법이다.

Lasso 계수들 $\hat{\beta}_\lambda^L$은 다음 식의 값을 최소로 만든다.

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = \text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j| \tag{6}$$

식 (6)과 (4)를 비교해보면 Lasso와 능형회귀는 비슷한 형태를 갖는다는 것을 알 수 있다. 유일한 차이는 (4)의 능형회귀 패널티에서 β_j^2항이 (6)의 Lasso 패널티에서는 $|\beta_j|$로 대체되었다는 것이다. 통계적 용어로 Lasso는 ℓ_2 패널티 대신 ℓ_1패널티를 사용한다. 계수 벡터 β의 ℓ_1-norm은 $||\beta||_1 = \sum |\beta_j|$로 주어진다.

능형회귀에서와 같이 Lasso는 계수 추정치들을 0으로 수축한다. 하지만 Lasso에서 ℓ_1 패널티는 조율 파라미터 λ가 충분히 클 경우 계수 추정치들의 일부를 정확히 0이 되게 하는 효과를 갖는다. 따라서, Lasso는 계수 수축뿐만 아니라 변수 선택도 수행하는셈이다. 그 결과 Lasso로부터 생성된 모형은 능형회귀에 의해 생성된 것보다 일반적으로 해석하기 훨씬 더 쉽다.

하나의 예로 Credit 자료에 Lasso를 적용하여 얻은 Figure 6.6의 계수 그래프를 살펴보자.

[4] Lasso was introduced by Robert Tibshirani in 1996 ....

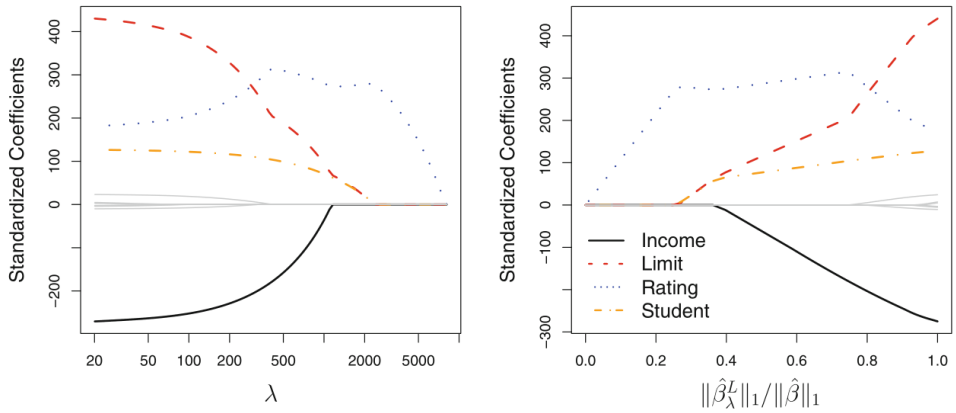


FIGURE 6.6. *The standardized lasso coefficients on the* `Credit` *data set are shown as a function of* λ *and* $\|\hat{\beta}_\lambda^L\|_1/\|\hat{\beta}\|_1$.

λ의 값에 따라 Lasso는 임의의 수의 변수들을 포함하는 모형을 생성할 수 있음을 확인할 수 있다.

능형회귀와 Lasso에 대한 또 다른 구성

Lasso와 능형회귀 계수 추정치는 각각 다음 문제를 푸는 것이라는 것을 보여줄 수 있다.

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s \tag{7}$$

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s \tag{8}$$

다르게 말하면, λ의 모든 값에 대해 식 (6)과 (7)은 동일한 Lasso 계수 추정치를 제공하는 어떤 s가 있다. 마찬가지로, λ의 모든 값에 대해 식 (4)와 (8)은 동일한 능형회귀 계수 추정치를 제공하는 어떤 s가 있다.

Figure 6.7는 $p = 2$일 때의 상황을 보여준다. 식 (7)은 Lasso 계수 추정치가 $|\beta_1| + |\beta_2| \leq s$에 의해 정의된 마름모 내의 모든 점 중에서 가장 작은 RSS를 갖는다는 것을 나타낸다.
마찬가지로, 능형회귀 추정치는 $\beta_1^2 + \beta_2^2 \leq s$로 정의된 원 내부에 있는 모든 점 중에서 가장 작은 RSS를 갖는다.

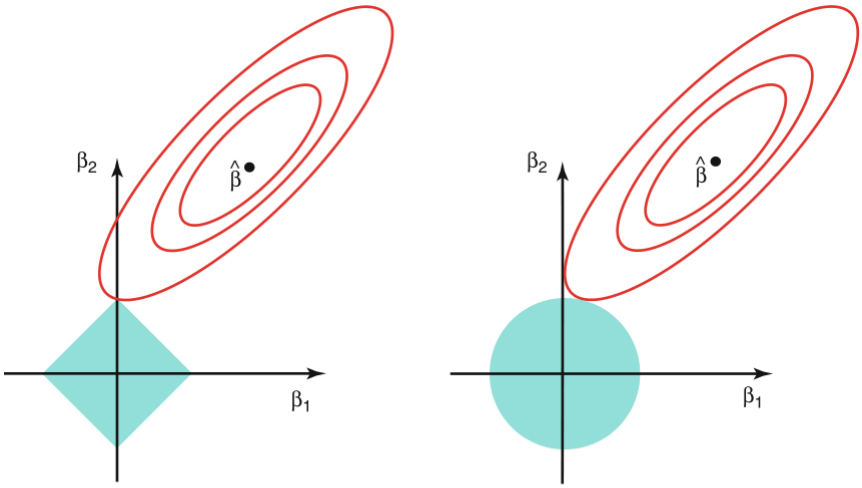
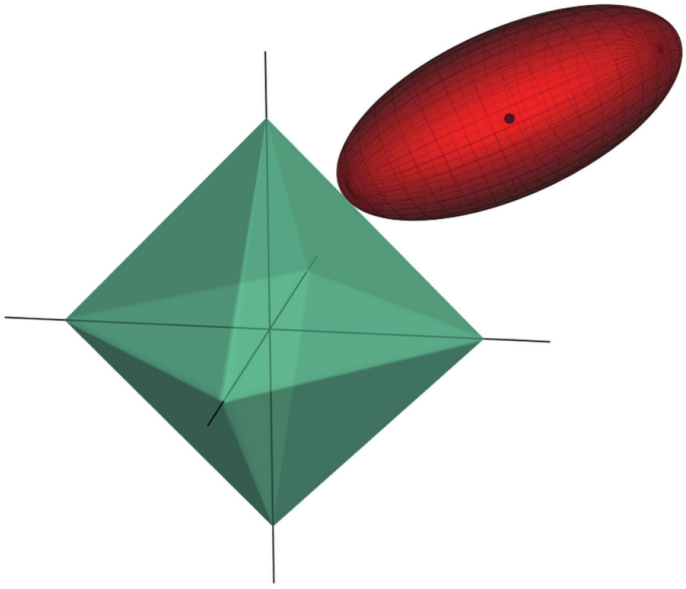


FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$, *while the red ellipses are the contours of the RSS.*



능형회귀는 뾰족한 부분이 없는 원형의 제한영역을 가지고 있어 RSS 타원과의 교점이 일반적으로 축상에 있지 않고, 그래서 능형회귀 계수 추정치는 0이 되지 않을 것이다. 하지만 Lasso 제한영역은 각 축에 모서리를 가지고 있어 RSS타원은 종종 축에서 제한영역과 만나게 될 것이다. 이렇게 되면 계수들 중 하나는 0이 될 것이다.

Lasso와 능형회귀 비교

Lasso는 설명변수들 중 일부만 포함하여 더 단순하고 해석력이 높은 모형을 생성한다. 이것이 능형회귀에 비해 Lasso가 갖는 주요 장점이다. 하지만 어느 방법이 더 낮은 Test MSE를 제공할까?

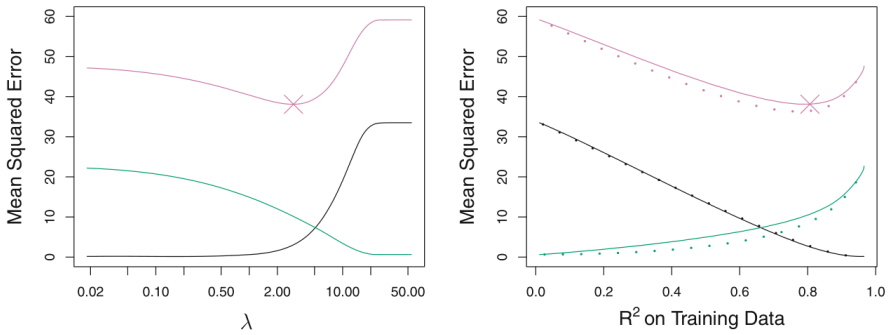


FIGURE 6.8. Left: *Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso on a simulated data set.* Right: *Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dotted). Both are plotted against their* R^2 *on the training data, as a common form of indexing. The crosses in both plots indicate the lasso model for which the MSE is smallest.*

Figure 6.8은 Figure 6.5와 동일한 모의 자료에 적용한 Lasso의 분산, 제곱편향, 그리고 Test MSE를 나타낸다. Figure 6.8의 오른쪽 패널 그래프는 Trainin data의 R^2에 대해 그린 것이다. 이것은 모형을 식별하는 또 다른 유용한 방법이고 서로 다른 유형의 정규화(규칙화; regularization)를 갖는 모형들을 비교하는 데 사용될 수 있다. 이 예제에서 Lasso와 능형회귀의 편향은 거의 동일하지만 능형회귀의 분산이 Lasso의 분산보다 약간 낮다. 따라서 능형회귀의 최소 MSE가 Lasso의 것보다 약간 작다.

하지만 Figure 6.8에서 사용된 데이터는 45개의 설명변수 모두가 반응변수와 관련되어 있는 방식으로 생성되었다. 즉, 실제 계수들 $\beta_1, \cdots, \beta_{45}$ 중 어느 것도 0이 아니다. Lasso는 암묵적으로 실제 계수들 중 다수가 0이라고 가정한다. 그 결과, 이 설정에서는 능형회귀가 Lasso보다 Test MSE측면에서 성능이 더 나은것이 놀랍지 않다. 아래 Figure 6.9는 유사한 상황을 보여주며, 다른 점은 반응변수가 45개의 설명변수 중 단 2개의 함수라는 것이다. 여기서는 Lasso가 편향, 분산, 그리고 MSE 측면에서 능형회귀보다 성능이 좋은 경향이 있다.

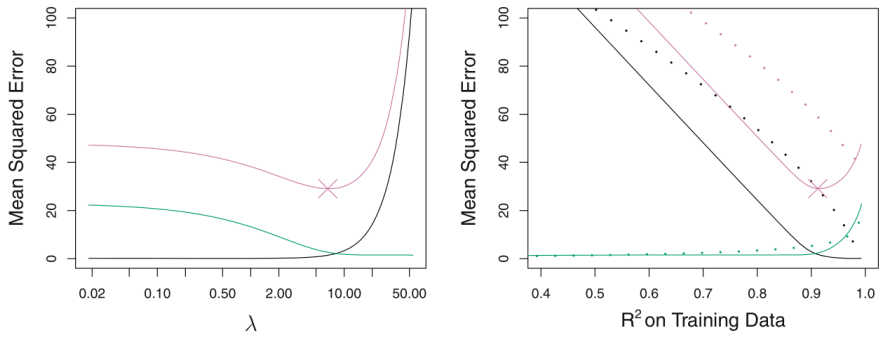


FIGURE 6.9. Left: *Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso. The simulated data is similar to that in Figure 6.8, except that now only two predictors are related to the response.* Right: *Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dotted). Both are plotted against their* R^2 *on the training data, as a common form of indexing. The crosses in both plots indicate the lasso model for which the MSE is smallest.*

이 두 예제는 능형회귀와 Lasso 중 어느 하나가 다른 것보다 항상 좋은 것은 아니라는 것을 보여준다.

일반적으로 Lasso는 비교적 적은 수의 설명변수가 상당히 큰 계수를 가지고 나머지 변수들은 계수가 아주 작거나 0인 설정에서 성능이 더 나을 것이라 기대할 수 있을 것이다.

능형회귀는 반응변수가 많은 설명변수들의 함수이고 그 계수들이 거의 동일한 크기일 때 성능이 더 좋을 것이다.

능형회귀와 Lasso에 대한 특별한 사례

능형회귀와 Lasso의 동작에 대해 더 나은 직관을 얻기 위해 $n = p$이고 대각원소는 1(비대각원소들은 0)인 대각행렬 $\mathbf{X}$를 갖는 단순한 경우를 고려해보자. 문제를 더 간단히 하기 위해 절편이 없는 회귀를 수행한다고 가정하자. 이러한 가정들로 인해 보통의 최소제곱 문제는 다음 식을 최소화하는 $\beta_1, \cdots, \beta_p$를 찾는 문제로 단순화된다.

$$\sum_{j=1}^{p} (y_j - \beta_j)^2 \tag{9}$$

이 경우에, 최소제곱 해는 다음과 같이 주어진다.

$$\hat{\beta}_j = y_j$$

이 설정에서 능형회귀는 다음 식 (10)을 최소로 하는 $\beta_1, \cdots, \beta_p$를 찾는 것이고,

$$\sum_{j=1}^{p} (y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{10}$$

Lasso는 식 (11)을 최소로 하는 계수들을 찾는 것이다.

$$\sum_{j=1}^{p} (y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{11}$$

이 설정에서 능형회귀 추정치는 (12)의 형태를 가지고 Lasso 추정치는 (13)의 형태를 갖는다.

$$\hat{\beta}_j^R = \frac{y_j}{1 + \lambda} \tag{12}$$

$$\hat{\beta}_j^L = \begin{cases} y_j - \lambda/2, & y_j > \lambda/2 \text{ 인 경우} \\ y_j + \lambda/2, & y_j < -\lambda/2 \text{ 인 경우} \\ 0, & |y_j| \leq \lambda/2 \text{ 인 경우} \end{cases} \tag{13}$$

아래 Figure 6.10은 이 상황을 나타낸다.

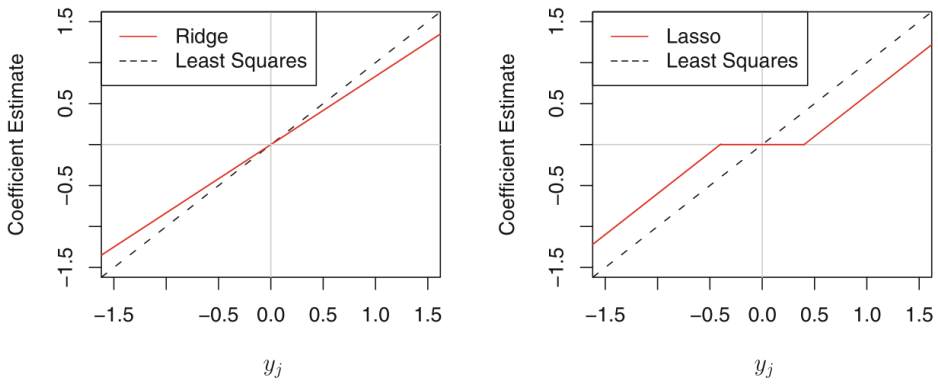


FIGURE 6.10. *The ridge regression and lasso coefficient estimates for a simple setting with $n = p$ and $\mathbf{X}$ a diagonal matrix with 1's on the diagonal.* Left: *The ridge regression coefficient estimates are shrunken proportionally towards zero, relative to the least squares estimates.* Right: *The lasso coefficient estimates are soft-thresholded towards zero.*

능형회귀와 Lasso는 매우 다른 형태의 수축(shrinkage)을 수행한다는 것을 볼 수 있다. 능형회귀는 각각의 최소제곱 계수 추정치를 같은 비율로 수축한다. 반면에 Lasso는 각각의 최소제곱 계수를 일정한 양 $\lambda/2$만큼 0으로 수축한다. 절대값이 $\lambda/2$보다 작은 최소제곱 추정치는 완전히 0으로 수축된다. 이 단순한 설정 (13)에서 Lasso에 의해 수행된 수축 유형은 소프트 임계처리(soft-thresholding)라고 알려져 있다. 일부 Lasso 계수들이 완전히 0으로 수축된다는 사실은 Lasso가 변수 선택을 수행하게 되는 이유를 설명해 준다.

좀 더 일반적인 데이터 행렬 $\mathbf{X}$의 경우에는 Figure 6.10에서 보여준 것보다 약간 더 복잡해지지만 주요 개념은 여전히 성립한다. 즉, 능형회귀는 데이터의 모든 차원을 같은 비율로 수축하는 반면, Lasso는 모든 계수들을 같은 양만큼 0을 향해 수축하고 충분히 작은 계수들은 완전히 0으로 수축한다.

6.2.3 조율 파라미터 선택(Selecting the Tuning Parameter)

능형회귀와 Lasso 모형을 적합하기 전에 먼저 조율 파라미터 λ값을 지정해 주어야하는데, 어떤 기준으로 λ값을 선택해야 할까? Test MSE를 최소로 만들어주는 λ값을 선택하면 될 것이다. 그런데 모집단의 모든 원소를 알고있는 것이 아니기 때문에 Test MSE를 추정해야 한다. 앞에서 학습한 LOOCV나 k-fold CV와 같은 직접 추정 방법을 이용해 Test MSE를 추정하고, 이를 토대로 λ값을 결정한다. 마지막으로, 이용 가능한 모든 관측치들과 선택된 조율 파라미터 값을 사용하여 모형을 다시 적합한다.

아래 Figure 6.12는 Credit 자료에 대한 능형회귀 적합에 LOOCV를 수행해서 얻은 결과로부터 λ를 선택하는 것을 나타낸다.

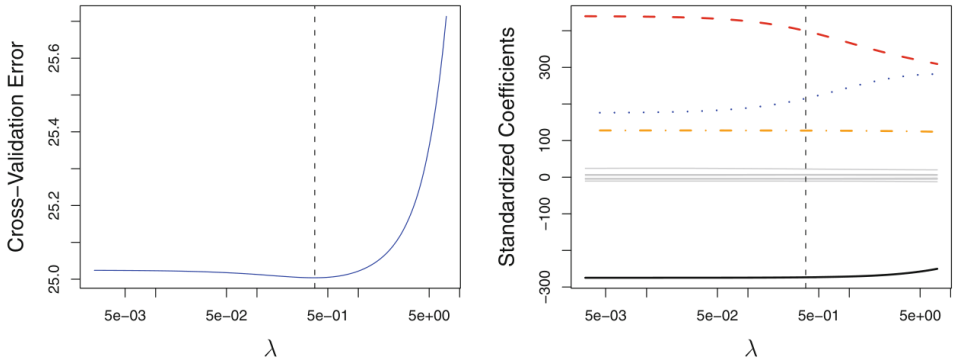


FIGURE 6.12. Left: *Cross-validation errors that result from applying ridge regression to the* `Credit` *data set with various value of* λ. Right: *The coefficient estimates as a function of* λ. *The vertical dashed lines indicate the value of* λ *selected by cross-validation.*

수직 점선은 선택된 λ의 값을 나타낸다. 이 경우에 값이 비교적 작은 것은 최적의 적합으로 얻는 최소제곱 해에 대한 수축량이 크지 않음을 의미한다. 게다가, 곡선에 깊이 내려간 부분이 없어 넓은 범위의 값들이 아주 비슷한 오차를 제공할 것이다. 이와 같은 경우에는 단순히 최소제곱 해를 이용할 수도 있다.

아래 Figure 6.13은 Figure 6.9의 스파스(sparse) 모의 자료에 대한 Lasso 적합에 10-fold 교차검증을 적용한 것이다.

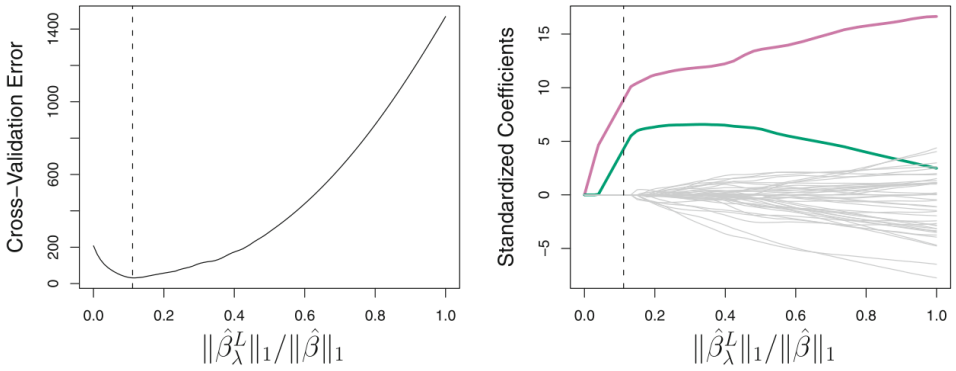


FIGURE 6.13. Left*: Ten-fold cross-validation MSE for the lasso, applied to the sparse simulated data set from Figure 6.9.* Right: *The corresponding lasso coefficient estimates are displayed. The vertical dashed lines indicate the lasso fit for which the cross-validation error is smallest.*

Figure 6.13의 오른쪽 패널에서 2가지 색의 선들은 반응변수와 관련이 있는 2개의 설명변수를 나타내고, 회색 선들은 관련없는 설명변수들을 나타낸다. 흔히 반응변수와 관련있는 설명변수를 **신호 (signal)** 변수라 하고 관련없는 변수는 **잡음 (noise)** 변수라고 한다. Lasso는 정확하게 2개의 신호변수에 훨씬 큰 계수 추정치를 제공한다. 뿐만 아니라 신호변수들만이 0이 아닌 계수 추정치를 갖게 해줬음을 알 수 있다. 그러므로 $p = 45$개의 변수와 $n = 50$관측치를 갖는 어려운

설정임에도 불구하고 Lasso와 10-fold 교차검증은 모형에서 2개의 신호변수들을 정확하게 식별한다. 반면에 Figure 6.13의 오른쪽 패널에서 오른쪽 끝 부분에 나타낸 최소제곱 해는 두 개의 신호변수 중 하나에만 큰 계수 추정치를 할당한다.