회귀분석팀

6팀
유종석
윤경선
채소연
김진혁
안은선

INDEX

1

다중공선성

1 다중공선성

다중공선성이란

다중공선성

모델에서 설명변수 X_j들 사이에 서로 **선형적인 관계**가 존재

변수에 대한 가정

- 선형성
- 설명변수들은 서로 독립
- 설명변수는 확률변수가 아님

다중공선성은 설명변수 간 독립적이어야 한다는 가정을 위배

1 다중공선성

다중공선성이란

다중공선성

모델에서 설명변수 X_j들 사이에 서로 **선형적인 관계**가 존재

변수에 대한 가정

선형성

설명변수들은
서로 독립

설명변수는
확률변수가 아님

다중공선성은 설명변수 간 독립적이어야 한다는 가정을 **위배**

1 다중공선성

다중공선성이란

예시

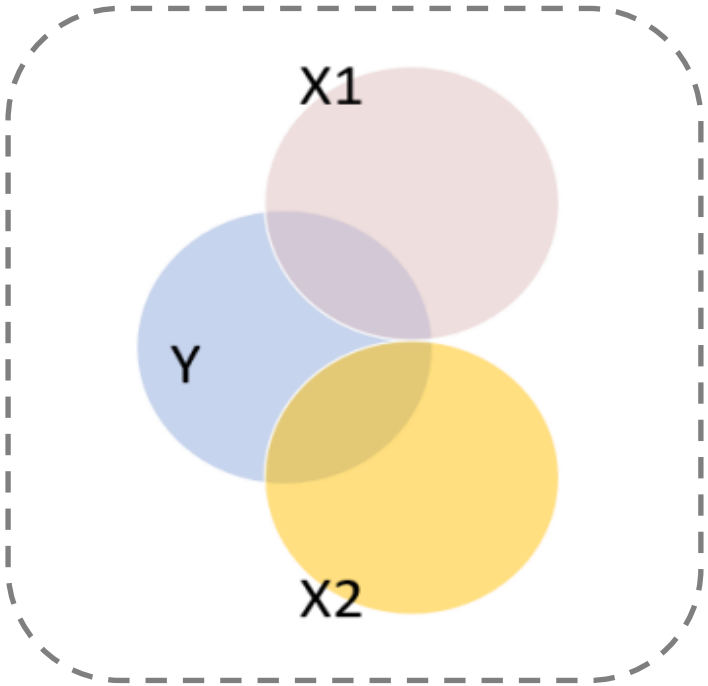
Y : 학점, X_1 : 오답 수, X_2 : 정답률, X_3 : 전체 문제 수

X_2 변수는 $X_2 = \left(1 - \frac{X_1}{X_3}\right) \times 100$으로 X_1, X_3 로 인해 완벽하게 설명됨

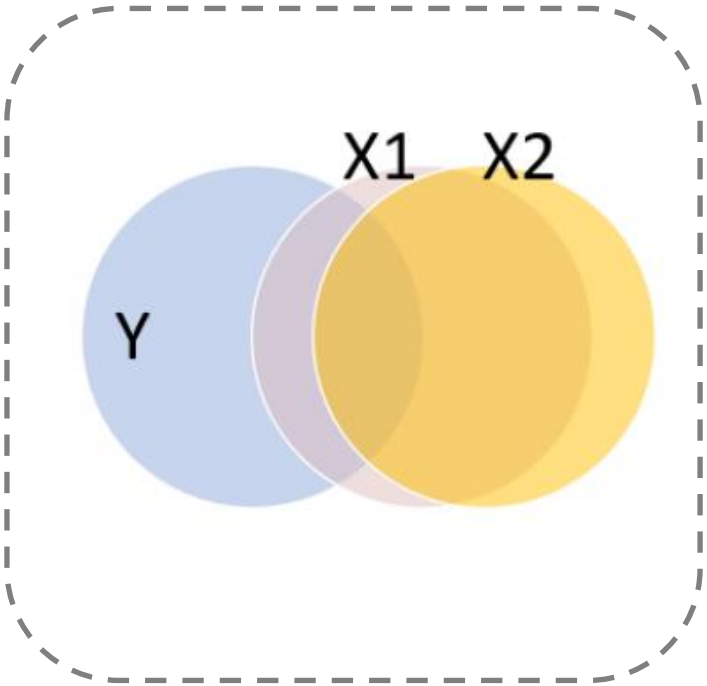
X_2 변수는 완전히 **필요하지 않은 변수**

1 다중공선성

다중공선성이란



✓ 다중공선성이 없는 경우



✓ 다중공선성이 있는 경우

다중공선성의 문제 | ① 추정량의 문제

추정량의 문제

① 모수의 추정 자체를 어렵게 만듦

최소제곱법을 통한 LSE로 적합된 다중선형회귀모형

$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'y$$

'역행렬'이 존재하기 위해서는?

$X'X$가 **full rank**이어야 함 → X가 **full-column rank** 행렬이 되어야 함 → X의 p개 변수가 **선형독립**이어야 함

full rank : 정방행렬 $X'X$의 모든 열 혹은 행이 선형독립이며 행렬식이 0이 아님

1 다중공선성

다중공선성의 문제 | ① 추정량의 문제

추정량의 문제

① 모수의 추정 자체를 어렵게 만듦

반대로 **선형종속**이라면 $X'X$의 **역행렬은 존재하지 않음 (다중공선성)**

최소제곱법(OLS method)을 사용할 수 없게 됨!

→ 모수의 추정 자체가 어려워짐

Complete Multicollinearity (완전한 선형종속)

현실에서는 근사적으로 선형종속을 이루는 경우가 빈번히 발생함!

다중공선성의 문제 | ① 추정량의 문제

추정량의 문제

② 추정량을 불안정하게 만듦

근사적으로 선형종속이 존재한다면 $\det(X'X) \approx 0$

$$Var(\hat{\beta}) = \sigma^2 (X'X)^{-1}$$

$$(X'X)^{-1} = \frac{1}{\det(X'X)} adj(X'X)$$

추정량의 분산도 급격하게 커져버려 **계수의 추정이 불안정**해진다는 문제 발생

다중공선성의 문제 | ② 모델의 문제

모델의 문제

① 모델의 검정 결과를 신뢰할 수 없음

전체 회귀식(F-test)은 유의한데,

개별 회귀계수 중에는 **유의한 것이 없는** 결과 발생

$Var(\hat{\beta}_j)$

추정량의 분산 커짐

$\hat{\beta}_j / \sqrt{Var(\hat{\beta}_j)}$

검정통계량 감소

(t-test)

귀무가설 기각 X

다중공선성의 문제 | ② 모델의 문제

모델의 문제

② 모델 해석에도 영향을 줌

개별 베타계수 β_j 의 해석

β_j

다중선형회귀모델

x_j 를 제외한 **나머지 변수가 고정되어 있을 때,**

x_j가 한 단계 증가하면 증가하게 되는 증가량

다중공선성의 문제 | ② 모델의 문제

모델의 문제

② 모델 해석에도 영향을 줌

다중공선성이 있는 경우 x_j가 변할 때

선형종속관계에 있는 다른 변수들도 **변할 수 있음**

개별 베타계수 β_j 의 해석

β_j

x_j 를 제외한 나머지 변수가 고정되어 있을 때,

x_j가 한 단계 증가하면 증가하게 되는 증가량

다중선형회귀모델

다중공선성이 존재하는 경우 개별 베타 계수 β_j를 해석할 때

'나머지 변수가 고정되어 있을 때'라는 가정 상황이 **불가능**

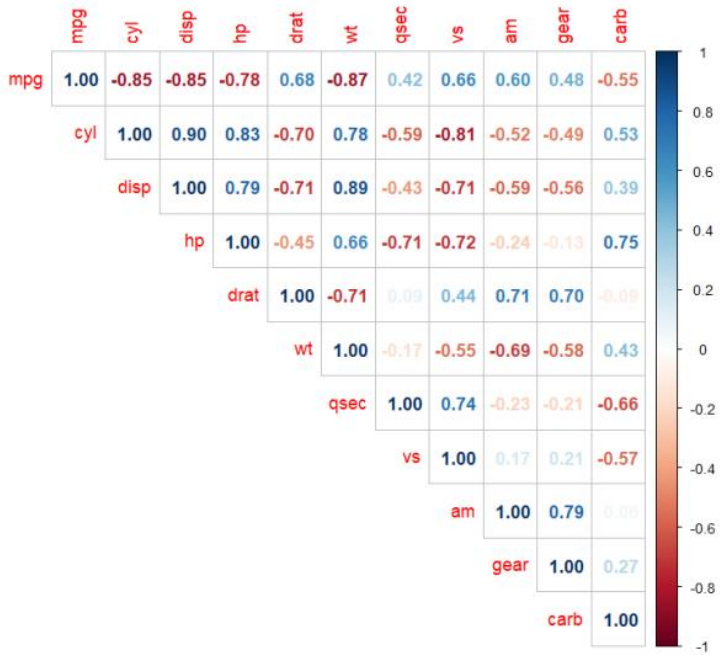
다중공선성의 진단 | ① 직관적인 판단

① F-test는 유의했지만 개별 회귀계수들에 대한 검정에서 귀무가설을 대부분을 기각하지 못할 경우

② 상식적으로 유의한 회귀계수가 유의하지 않다고 나올 경우

③ 추정된 회귀계수의 부호가 상식과 다를 경우

다중공선성의 진단 | ② 상관계수 plot



상관계수 plot을 통해...

- ✓ 변수들 사이의 선형관계 여부 파악 가능
- ✓ 보통 절댓값을 기준으로 상관계수가 0.7 이상일 경우 다중공선성 의심

다중공선성의 진단 | ③ VIF

VIF(분산팽창인자)

$$VIF_j = \frac{1}{1 - R_j^2}, \qquad j = 1, \dots, p$$

R_j^2

다중선형회귀모델 $x_j = \gamma_1 x_1 + \cdots + \gamma_{j-1} x_{j-1} + \gamma_{j+1} x_{j+1} + \cdots \gamma_p x_p$을 적합했을 때의 결정계수로, **회귀식이 데이터를 설명하는 정도**를 의미

R_j^2가 크다 → x_j가 **나머지 변수들의 선형결합**으로 충분히 표현 가능 → **다중공선성** 존재

다중공선성의 진단 | ③ VIF

VIF(분산팽창인자)

$$VIF_j = \frac{1}{1 - R_j^2}, \qquad j = 1, \dots, p$$

R_j^2

다중선형회귀모델 $x_j = \gamma_1 x_1 + \cdots + \gamma_{j-1} x_{j-1} + \gamma_{j+1} x_{j+1} + \cdots \gamma_p x_p$을 적합했을 때의 결정계수로, **회귀식이 데이터를 설명하는 정도**를 의미

- ✓ 일반적으로 VIF가 **10 이상일 경우** 심각한 다중공선성 의미
- ✓ 다중공선성이 완전히 없다면 VIF 값은 1

1 다중공선성

다중공선성의 해결

① 변수선택법
(Variable Selection)

② 차원축소
(Dimension Reduction)

선대팀 3주차 클린업 참고~

③ 정규화 (Regularization, Penalizing, Shrinkage)

④ 필터링 방법

- ✓ 차원축소 방법에는 PCA, PLS, 신경망 모형을 사용하는 AE, 요인분석 등이 있음
- ✓ 필터링 방법은 모델링 전 변수 자체의 통계적인 특징만으로 변수를 선택하는 방법

2

변수선택법

2 변수선택법

변수선택법이란?

분석을 위해 고려한 수많은 변수들 중 **적절한 변수의 조합**을 찾아내는 방법

우리에게 주어진 가능한 후보 변수(Candidate Regressor)들 중
일부분만 중요하거나 **예측에 유의미**할 수 있음

후보 변수들의 적절한 부분집합(subset)을 찾아 다중공선성 해결

2 변수선택법

변수선택법이란?

분석을 위해 고려한 수많은 변수들 중 **적절한 변수의 조합**을 찾아내는 방법

우리에게 주어진 가능한 후보 변수(Candidate Regressor)들 중
일부분만 중요하거나 **예측에 유의미**할 수 있음

후보 변수들의 적절한 부분집합(subset)을 찾아 다중공선성 해결

2 변수선택법

변수선택법이란?

변수가 선택되고 제거되는 것에 논리성과 정당성을 부여해주는 방법

분석을 위해 고려한 수많은 변수들 중 적절한 변수의 조합을 찾아내는 방법

변수선택법을 통해 다중공선성이 완벽히 제거되지는 않지만,
높은 상관관계를 가지는 변수들 중 일부만을 선택할 수 있고,
변수의 선택과 제거에 대한 정당성 부여 가능

다중공선성이 발견되지 않더라도
최종 모델에 대한 해석력 증가 등
확신을 얻을 목적으로 시행하기도 함

2 변수선택법

변수선택법이란?

최대한 많은 변수들을 사용해서

y를 예측하기 위한 많은 정보를 포함하고 싶기도 하지만,

최대한 적은 변수들을 사용해서 모형의 분산을 줄이고도 싶어함.

최적의 회귀식(Best Regression Equation)을 heuristic하게 찾는 방법

"변수선택법"

변수가 너무 적으면 간결하고 해석이 쉬워지더라도 예측력은 떨어질 것이고,

변수가 너무 많으면 과적합(Overfitting)될 가능성이 높음

2 변수선택법

변수 선택 지표

1주차 클린업에서...

Partial F-test

유의하지 않은 변수들을 없애는 방식으로 변수 선택을 진행할 수 있음

Full Model(FM)과 Reduced Model(RM)이 서로 내포(nested)관계에 있어야 함

= RM에 있는 모든 변수가 FM에 있어야 함

$$modelA: y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad \text{VS} \quad modelB: y = \beta_0' + \beta_3 x_3 + \beta_4 x_4$$

변수 선택 지표

내포 관계에 있지 않은 여러 모델들을 만들어 비교해야 하는 경우 존재

$$modelA : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad \text{VS} \quad modelB : y = \beta_0' + \beta_3 x_3 + \beta_4 x_4$$

$modelA$ 와 $modelB$는 사용된 변수가 전혀 달라 내포 관계(nested)가 아님

일반적인 상황에서도 (포함관계와 무관하게)

모델의 설명력과 변수의 개수를 모두 고려해주는 지표 필요

2 변수선택법

변수 선택 지표

내포 관계에 있지 않은 여러 모델들을 만들어 비교해야 하는 경우 존재

$$modelA : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad \text{VS} \quad modelB : y = \beta_0' + \beta_3 x_3 + \beta_4 x_4$$

$modelA$ 와 $modelB$는 사용된 변수가 전혀 달라 내포 관계(nested)가 아님

일반적인 상황에서도 (포함관계와 무관하게)

모델의 설명력과 변수의 개수를 모두 고려해주는 지표 필요

2 변수선택법

변수 선택 지표

Check!

- ✓ 모델의 설명력
- ✓ 변수의 개수

내포 관계에 있지 않은 여러 모델들의 적합도를 비교해야 하는 경우 존재

$modelA : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ vs $modelB : y = \beta_0' + \beta_3 x_3 + \beta_4 x_4$

$modelA$와 $modelB$는 사용된 변수가 전혀 달라 내포 관계(nested)가 아님

일반적인 상황에서도 (포함관계와 무관하게)

모델의 설명력과 **변수의 개수**를 **모두 고려**해주는 지표 필요

변수 선택 지표 | ① 수정결정계수

수정결정계수(R^2_{adj})

설명력을 담당하는 **결정계수**와 **변수 개수 패널티**가 들어감

👉 복합적 고려 가능

수정결정계수 식

$$R^2_{adj} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

변수 선택 지표 | ② AIC

Akaike Information Criterion

일반적인 계산식

$$AIC = -2\log(Likelihood) + 2p$$

$$AIC = nlog(2\pi\hat{\sigma}^2) + \frac{SSE}{\hat{\sigma}^2} + 2p$$

정규분포 가정 하의 AIC

p : 모델의 모수 개수

$\hat{\sigma}^2$: σ^2의 MLE

$Likelihood$ 가 **커질수록** 모델이 데이터를 **잘 설명**한다

변수 선택 지표 | ② AIC

Akaike Information Criterion

일반적인 계산식

$$AIC = -2\log(Likelihood) + 2p$$

$$AIC = n\log(2\pi\hat{\sigma}^2) + \frac{SSE}{\hat{\sigma}^2} + 2p$$

정규분포 가정 하의 AIC

p : 모델의 모수 개수

$\hat{\sigma}^2$: σ^2의 MLE

변수에 **개수**에 따라 **패널티 부과**

변수 선택 지표 | ② AIC

Akaike Information Criterion

일반적인 계산식

$$AIC = -2\log(Likelihood) + 2p$$

$$AIC = nlog(2\pi\hat{\sigma}^2) + \frac{SSE}{\hat{\sigma}^2} + 2p$$

정규분포 가정 하의 AIC

p : 모델의 모수 개수

$\hat{\sigma}^2$: σ^2의 MLE

$Likelihood$ 가 **커지면** AIC는 **작아지기** 때문에,

AIC가 **낮을수록 더 좋은 모형**이라고 해석

변수 선택 지표 | ③ BIC

Bayesian Information Criterion

일반적인 계산식

$$BIC = -2\log(Likelihood) + p \times log(n)$$

$$BIC = nlog(2\pi\hat{\sigma}^2) + \frac{SSE}{\hat{\sigma}^2} + p \times log(n)$$

정규분포 가정 하의 BIC

p : 모델의 모수 개수

$\hat{\sigma}^2$: σ^2의 MLE

AIC와 같이 **작을수록** 더 **좋은 모형**

변수 선택 지표 | ③ BIC

Bayesian Information Criterion

일반적인 계산식

$$BIC = -2\log(Likelihood) + p \times log(n)$$

$$BIC = nlog(2\pi\hat{\sigma}^2) + \frac{SSE}{\hat{\sigma}^2} + p \times log(n)$$

정규분포 가정 하의 BIC

p : 모델의 모수 개수

$\hat{\sigma}^2$: σ^2의 MLE

데이터의 개수를 **모수의 개수**에 곱함으로서

AIC보다 **더 큰 패널티** 부과

변수 선택 지표 | ③ BIC

Bayesian Information Criterion

일반적인 계산식

$$BIC = -2\log(Likelihood) + p \times log(n)$$

$$BIC = nlog(2\pi\hat{\sigma}^2) + \frac{SSE}{\hat{\sigma}^2} + p \times log(n)$$

정규분포 가정 하의 BIC

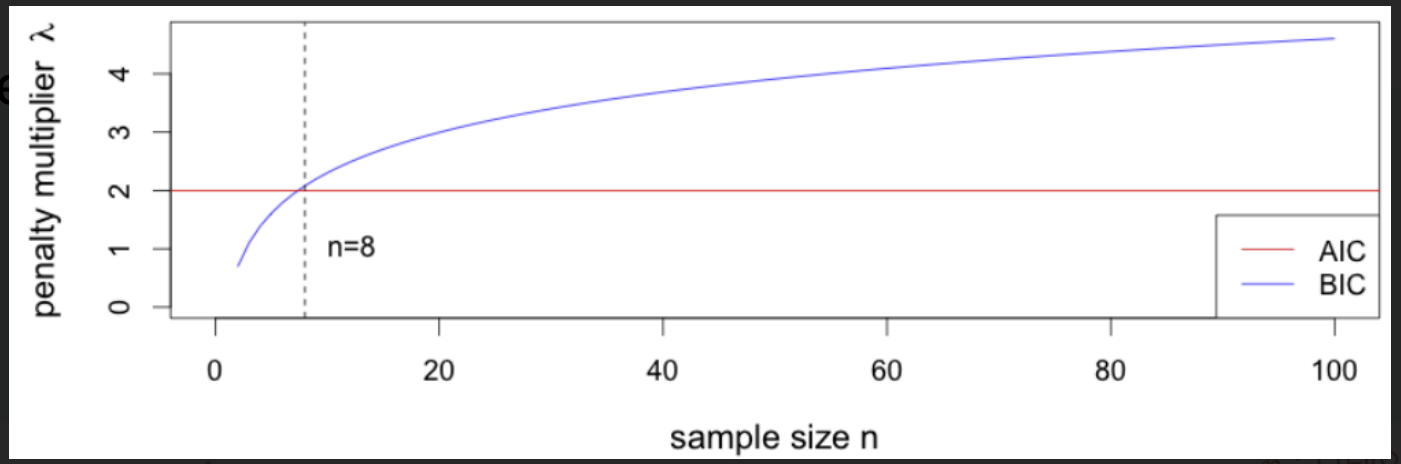
p : 모델의 모수 개수

$\hat{\sigma}^2$: σ^2의 MLE

n > 8이라면 BIC가 AIC보다 **더 많은 패널티**를 **부여**하기 때문에

변수 개수가 더 **적은 모델**을 **선호**

변수 선택 지표 | ③ BIC



➡ BIC가 AIC보다 변수 증가에 더 민감하므로,

변수의 개수가 작은 것이 우선순위라면 BIC를 참고하는 것이 좋음

n > 8이라면 BIC가 AIC보다 더 많은 패널티를 부여하기 때문에

변수 개수가 더 적은 모델을 선호

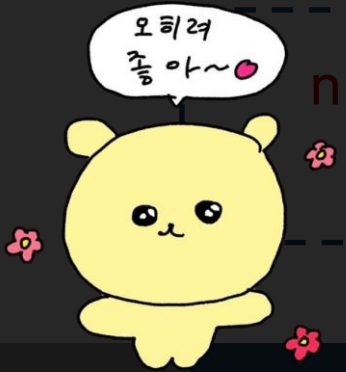
2 변수선택법

변수 선택 지표 | ③ BIC

Bayesian Information Criterion

$$BIC = -2\log(Likelihood) + p \times log(n)$$

고차원 데이터에서는 정확성이 떨어질 수 있고, AIC와 BIC 모두에서 문제가 발생할 수 있기에 종합적으로 고려해서 모형을 선택해야 함



n > 8이라면 BIC가 AIC보다 더 많은 패널티를 부여하기 때문에 변수 개수가 더 적은 모델을 선호

2 변수선택법

변수 선택 방법

변수선택법은 모두 heuristic(경험적인)한 방법

HOW

직접 알고리즘에 따라 해당하는 모든 경우를 계산해서

제일 좋은 회귀식을 찾는 방법

'직접'이라는 단어에 트라우마 있으신 분? 일단 저요

변수 선택 방법 | ① Best Subset Selection

Best Subset Selection

가능한 **모든 변수들**의 **조합**을 **다 고려**하는 방법

변수의 개수가 p 개라면, 2^p개의 모형을 모두 적합하고 비교

가능한 모든 경우의 수를 고려하기에
선택된 Best Model에 대한
더 신뢰할 수 있는 결과 산출

✓ 변수의 개수가 p > 40인 경우 계산 불가능

✓ 적당한 p에서도 많은 관측치를 지닐 경우
모든 모델을 고려한 계산 비용 많이 소모

2 변수선택법

변수 선택 방법 | ① Best Subset Selection

Best Subset Selection

가능한 모든 변수들의 조합을 만들어 비교하는 방법

변수의 개수가 p 개라면, 2^p개의 모형을 모두 적합하고 비교

Best Subset Selection's Algorithm

1. $M_1, \cdots, M_p$ 개의 모형을 적합한다. 이때 M_k(k = 1, 2, ⋯ p)란 변수의 개수를 k개로 적합했을 때 적합한 회귀식 중 training error(주로 MSE)가 제일 작은 식이다.

2. $(M_1 \sim M_p)$ p개의 모형 중 AIC 또는 BIC가 가장 작은 모형을 선택한다.

3. 만약 AIC, BIC가 가장 작은 모형이 서로 다를 경우 다른 근거에 의해 두 개의 모형 중 하나를 선택한다.(주로 하나의 평가 기준을 두고 선택합니다)

M_k 란 변수의 개수를 k개로 적합했을 때 적합한 회귀식 중 MSE가 제일 작은 식

변수 선택 방법 | ① Best Subset Selection

Best Subset Selection

가능한 **모든 변수들**의 **조합**을 **다 고려**하는 방법

변수의 개수가 p 개라면, 2^p개의 모형을 모두 적합하고 비교

가능한 모든 경우의 수를 고려하기에

선택된 Best Model에 대한

더 신뢰할 수 있는 결과 산출

✓ 변수의 개수가 p > 40인 경우 계산 불가능

✓ 적당한 p에서도 많은 관측치를 지닐 경우 모든 모델을 고려한 계산 비용 많이 소모

변수 선택 방법 | ① Best Subset Selection

Best Subset Selection

가능한 **모든 변수들**의 **조합**을 **다 고려**하는 방법

변수의 개수가 p 개라면, 2^p개의 모형을 모두 적합하고 비교

가능한 모든 경우의 수를 고려하기에
선택된 Best Model에 대한
더 신뢰할 수 있는 결과 산출

- ✓ 변수의 개수가 p > 40인 경우 계산 불가능
- ✓ 적당한 p에서도 많은 관측치를 지닐 경우
  모든 모델을 고려한 계산 비용 많이 소모

변수 선택 방법 | ② 전진선택법

전진선택법 *Forward Selection*

Null Model($y = \beta_0$)에서 시작해 변수를 하나씩 추가하는 방법

- Best Subset Selection에 비해 계산이 비교적 매우 빠름
- 변수의 개수가 관측치의 개수보다 많은 경우에도 사용 가능

Best Subset Selection처럼 가능한 모든 변수 조합을 고려하지는 않기 때문에 선택된 모형이 최적의 모형이라고 할 수 없음

2 변수선택법

변수 선택 방법 | ② 전진선택법

전진선택법 *Forward Selection*

Forward Selection's Algorithm

1. 상수항만을 포함하고 있는 모형인 Null Model($y = \beta_0$)에서 시작해 X_1부터 X_p까지의 변수들 중에 어떤 것을 추가하는 것이 AIC와 BIC를 낮추는지 판단한다.

2. 만약 1번의 과정에서 X_1이 선택되었다면, 이제 $y = \beta_0 + \beta_1 x_1$의 식에서 X_2부터 X_p까지의 변수들 중에 어떤 것을 추가하는 것이 AIC와 BIC를 낮추는지 판단하다.

3. 이러한 과정을 반복하며 AIC와 BIC가 낮아지면 추가하고 더 이상 AIC와 BIC가 낮아지지 않는다면 프로세스를 중단한다.

변수 선택 방법 | ② 전진선택법

전진선택법 *Forward Selection*

Null Model($y = \beta_0$)에서 시작해 변수를 하나씩 추가하는 방법

- Best Subset Selection에 비해 계산이 비교적 매우 빠름
- 변수의 개수가 관측치의 개수보다 많은 경우에도 사용 가능

Best Subset Selection처럼
가능한 모든 변수 조합을
고려하지는 않기 때문에
선택된 모형이 최적의 모형이라고 할 수 없음

변수 선택 방법 | ② 전진선택법

전진선택법 *Forward Selection*

Null Model($y = \beta_0$)에서 시작해 변수를 하나씩 추가하는 방법

- Best Subset Selection에 비해 계산이 비교적 매우 빠름
- 변수의 개수가 관측치의 개수보다 많은 경우에도 사용 가능

Best Subset Selection처럼
가능한 모든 변수 조합을
고려하지는 않기 때문에
선택된 모형이 최적의 모형이라고 할 수 없음

변수 선택 방법 | ③ 후진선택법

후진선택법 *Backward Elimination*

Full Model($y = \beta_0 + \beta_1 x_1 + \beta_1 x_2 + \cdots + \beta_p x_p$)에서 시작해

변수를 하나씩 제거하는 방법

Forward selection의 반대

Best Subset Selection에 비해
계산이 매우 빠름

✓ Best Subset Selection 방법과
마찬가지로 p>40인 경우에는 사용 X

✓ 가능한 모든 변수 조합을 고려하는 것은
아니기 때문에 선택된 모형이
최적의 모형이라고 할 수 없음

2 변수선택법

변수 선택 방법 | ③ 후진선택법

후진선택법 Backward Elimination

Full Model($Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$)에서 시작해

Backward Elimination's Algorithm

변수를 하나씩 제거하는 방법

1. Full Model에서 시작해 X_1부터 X_p 까지의 변수들 중에 가장 AIC와 BIC를 크게 낮추는 변수를 선택해 제거한다.

2. 위의 가정을 반복하며 AIC와 BIC가 더 이상 낮아지지 않으면 프로세스를 중단한다.

변수 선택 방법 | ③ 후진선택법

후진선택법 *Backward Elimination*

Full Model($y = \beta_0 + \beta_1 x_1 + \beta_1 x_2 + \cdots + \beta_p x_p$)에서 시작해

변수를 하나씩 제거하는 방법

Forward selection의 반대

Best Subset Selection에 비해
계산이 매우 빠름

✓ Best Subset Selection 방법과
마찬가지로 p>40인 경우에는 사용 X

✓ 가능한 모든 변수 조합을 고려하는 것은
아니기 때문에 선택된 모형이
최적의 모형이라고 할 수 없음

변수 선택 방법 | ③ 후진선택법

후진선택법 *Backward Elimination*

Full Model($y = \beta_0 + \beta_1 x_1 + \beta_1 x_2 + \cdots + \beta_p x_p$)에서 시작해

변수를 하나씩 제거하는 방법

Forward selection의 반대

Best Subset Selection에 비해

계산이 매우 빠름

- ✓ Best Subset Selection 방법과 마찬가지로 p>40인 경우에는 사용 X
- ✓ 가능한 모든 변수 조합을 고려하는 것은 아니기 때문에 선택된 모형이 최적의 모형이라고 할 수 없음

변수 선택 방법 | ④ 단계적 선택법

단계적 선택법 *Stepwise Elimination*

Forward Selection과 Backward Elimination 과정을 **섞은 방법**

Null model에서 시작할 수도 있고 Full Model에서 시작할 수도 있지만,
변수를 선택하거나 제거하는 경우를 모두 고려(조합)했을 때
AIC와 BIC가 **감소하는 방향으로 움직임**

2 변수선택법

변수 선택 방법 | ④ 단계적 선택법

단계적 선택법 *Stepwise Elimination*

Stepwise Selection's Algorithm 기본형

1. Null model 혹은 Full model에서 시작한다.

2. 다른 변수 선택법들을 혼합하여 변수들을 제거 혹은 추가하여 모델을 평가한다.

3. AIC/BIC가 가장 작은 모형을 선택한다.

Null model에서 시작할 수도 있고 Full Model에서 시작할 수도 있지만,

변수를 선택하거나 제거하는 경우를 모두 고려(조합)했을 때

AIC와 BIC가 감소하는 방향으로 움직임

2 변수선택법

변수 선택 방법 | ④ 단계적 선택법

단계적 선택법 Stepwise Elimination

Stepwise Selection's Algorithm 예시

1. 먼저 forward selection 과정을 이용해 가장 유의한 변수들을 모델에 추가한다

2. 그 후 나머지 변수들에 대해 Backward Elimination을 적용해 새롭게 유의하지 않게 된 변수들을 제거한다.

Null model에서 시작할 수도 있고 Full Model에서 시작할 수도 있지만,

3. 제거된 변수는 다시 모형에 포함되지 않으며, 모형에 유의하지 않은 설명변수가 존재하지 않을 때까지 1번과 2번 과정을 반복한다.

변수 선택 방법 | ④ 단계적 선택법

단계적 선택법 Stepwise Elimination

Forward Selection Backward Elimination 위 두 방법을 합쳐 놓은 방법

Stepwise Selection's Algorithm 예시

전진선택법을 사용할 때 한 변수가 선택되면,

이미 선택된 변수 중 중요하지 않은 변수가 있을 수 있다.

Null model에서 시작할 수도 있고 Full Model에서 시작할 수도 있지만,

전진선택법의 각 단계에서 이미 선택된 변수들의 중요도를

다시 검사하여 중요하지 않은 변수를 제거하는 방법

변수 선택 방법 | ④ 단계적 선택법

단계적 선택법 *Stepwise Elimination*

Forward Selection과 Backward Elimination 과정을 **섞은 방법**

Best Subset Selection에 비해
계산이 매우 빠름

변수를 제거 혹은 추가 모두를 할 수
있다는 점에서 유연하게 움직일 수 있지만,
모든 변수 조합을 고려하는 것이 아니기
때문에 Best Model이라고 할 수는 없음

변수 선택 방법 | ④ 단계적 선택법

단계적 선택법 *Stepwise Elimination*

Forward Selection과 Backward Elimination 과정을 **섞은 방법**

Best Subset Selection에 비해
계산이 매우 빠름

변수를 제거 혹은 추가 모두를 할 수
있다는 점에서 유연하게 움직일 수 있지만,
모든 변수 조합을 고려하는 것이 아니기
때문에 Best Model이라고 할 수는 없음

2 변수선택법

변수 선택 방법 | ④ 단계적 선택법

단계적 선택법 *Stepwise Elimination*

정리

Forward Selection과 Backward Elimination 과정을 **섞은 방법**

- Best Subset Selection을 제외한 나머지 방법들의 장점이 계산이 **매우 빠른 것**이라고 했지만, 이는 **상대적인 것**

- 위 방법 모두 **계산 비용**이 굉장히 **많이 소모**

- Forward Selection과 Backward Elimination의 결과를 고려했을 때, **둘의 결과가 상이할 수 있음**

변수를 제거 혹은 추가 모두를 할 수
있는 점에서 좋아보일 수 있지만,
모든 변수 조합을 고려하는 것이 아니기
때문에 Best Model이라고 할 수는 없음

2 변수선택법

변수 선택 방법 | ④ 단계적 선택법

단계적 선택법 *Stepwise Elimination*

정리

Forward Selection과 Backward Elimination 과정을 섞은 방법

기계적으로 변수를 추가 혹은 제거하는 행위는 매우 위험

Best Subset Selection에 비해
계산이 매우 빠름

정규화 방법

변수를 제거 혹은 추가 모두를 할 수
있다는 점에서 유연하게 움직일 수 있지만,
모든 변수 조합을 고려하는 것이 아니기
때문에 Best Model이라고 할 수는 없음

2 변수선택법

변수 선택 방법 | ④ 단계적 선택법

정리

단계적 선택법 *Stepwise Elimination*

Forward Selection과 Backward Elimination 과정을 섞은 방법

기계적으로 변수를 추가 혹은 제거하는 행위는 매우 위험

Best Subset Selection에 비해 계산이 매우 빠름

정규화 방법

변수를 제거 혹은 추가 모두를 할 수 있다는 점에서 유연하게 움직일 수 있고 모든 변수 조합을 고려하는 것이 아니기 때문에 Best Model이라고 할 수

3

정규화

정규화

정규화 *Regularization*

회귀계수가 가질 수 있는 값에 **제약조건**을 부여함으로써

계수들을 작게 만들거나 0으로 만드는 방법

다중공선성은 OLS 추정량의 분산을 크게 증가시킴

정규화는 OLS 추정량의 **불편성 포기** → **분산을 줄이는 효과**가 있음

정규화

정규화 *Regularization*

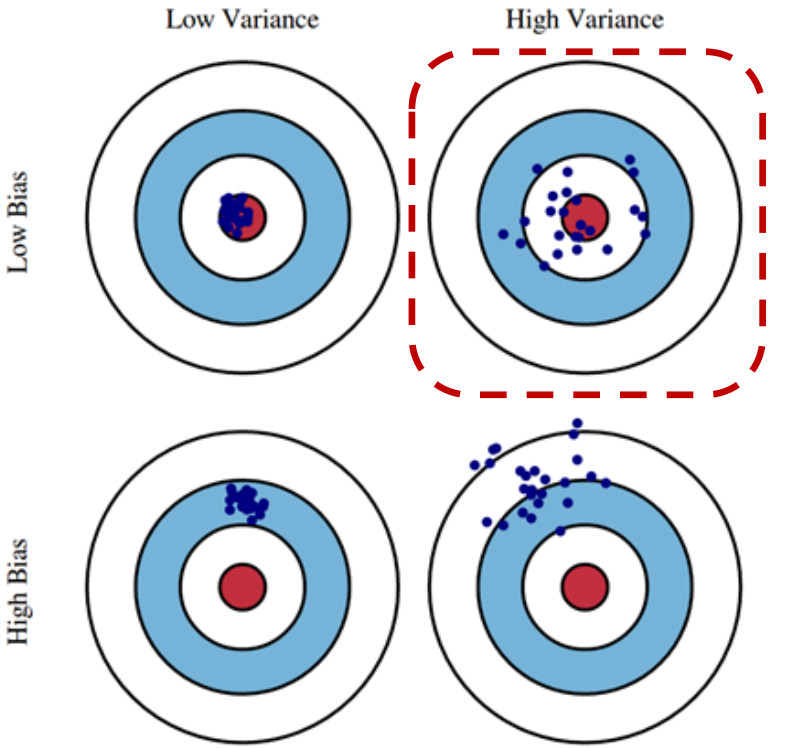
회귀계수가 가질 수 있는 값에 **제약조건**을 부여함으로써

계수들을 작게 만들거나 0으로 만드는 방법

다중공선성은 OLS 추정량의 분산을 크게 증가시킴

정규화는 OLS 추정량의 **불편성 포기** → **분산을 줄이는 효과**가 있음

정규화 | Bias-Variance Trade off



정규화 이전

다중공선성이 존재하는 경우

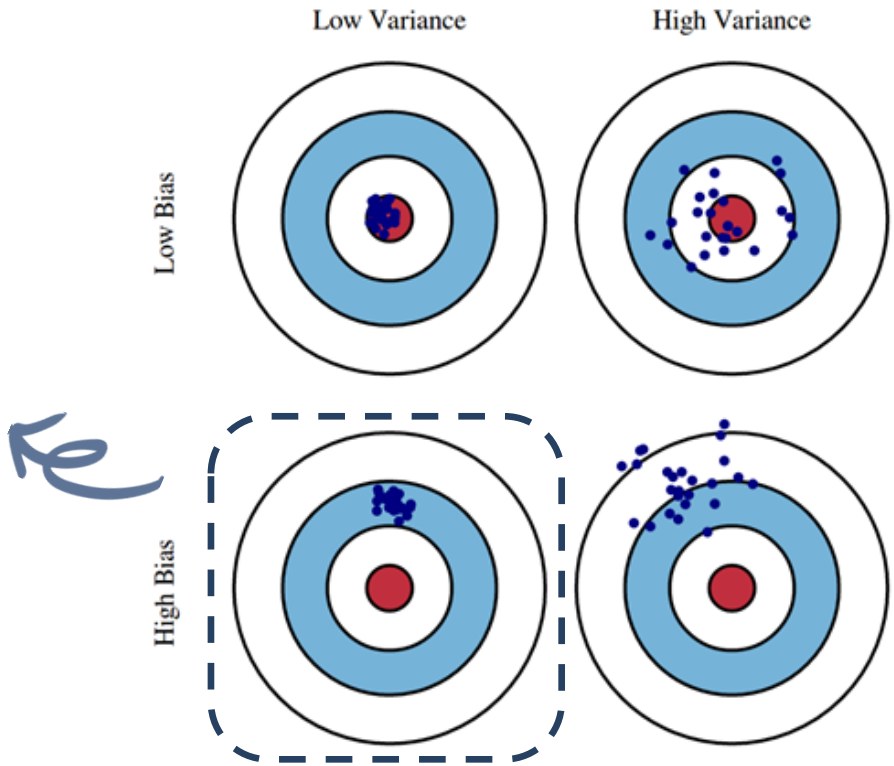
회귀계수들의 OLS 추정량

정규화 | Bias-Variance Trade off

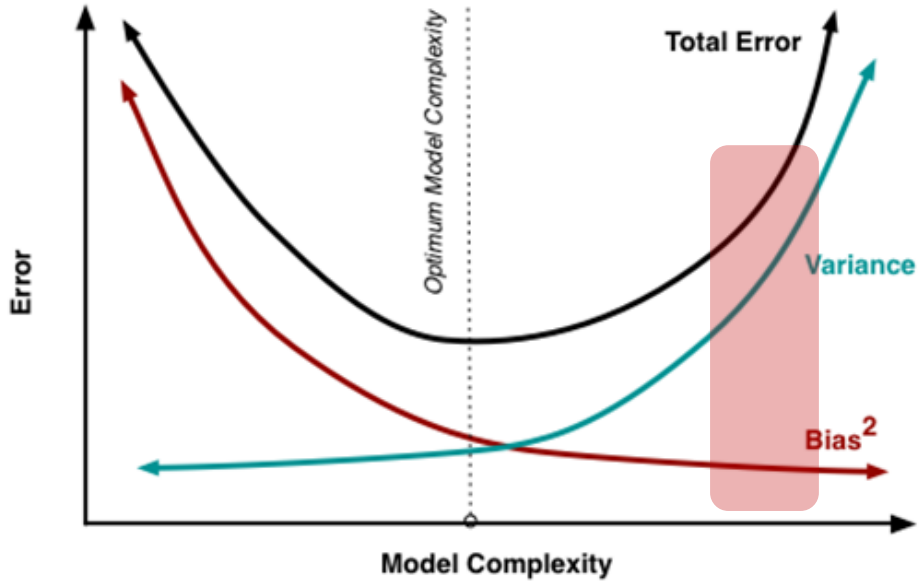
정규화 이후

정규화로 불편성을 포기하고

분산을 줄이는 효과



정규화 | Bias-Variance Trade off



정규화 이전

다중공선성이 존재하는 경우

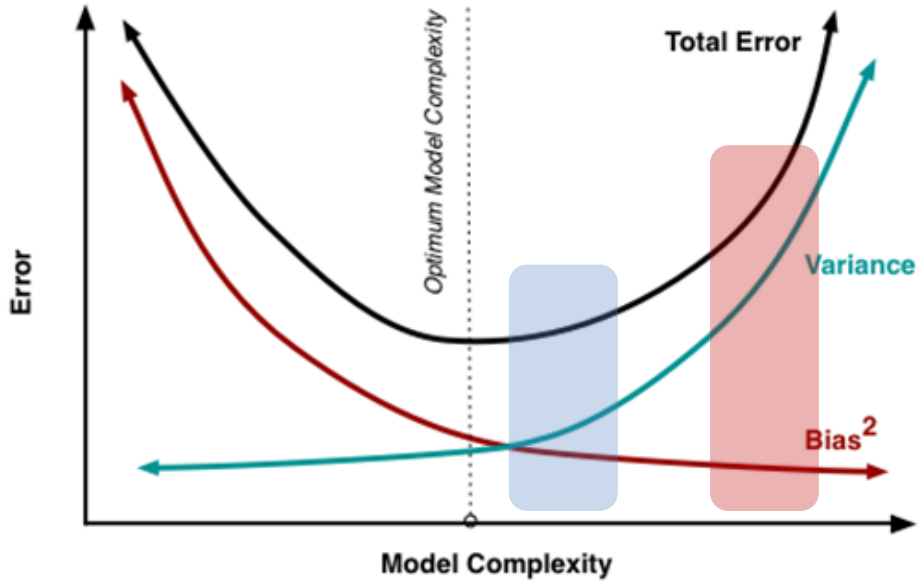
Low Bias & High Variance

정규화 이후

편향은 증가하지만

분산이 획기적으로 줄어듦

정규화 | Bias-Variance Trade off



정규화 이전

다중공선성이 존재하는 경우

Low Bias & High Variance

정규화 이후

편향은 증가하지만

분산이 획기적으로 줄어듦

정규화 | Bias-Variance Trade off

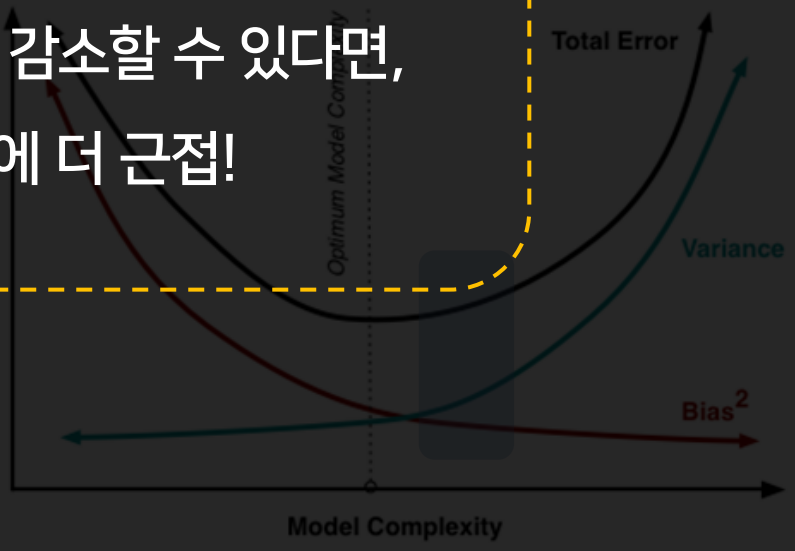
정규화 이후

편향은 증가하지만 분산이 획기적으로 줄어듦

bias의 증가폭 < variance의 감소폭

expected MSE가 감소할 수 있다면,

최적의 모델에 더 근접!



정규화

목적함수

$$L(\beta) = \min_{\beta} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Training Accuracy에 해당하는 LSE

Generalization Accuracy (정규화를 진행했다는 증거)

3 정규화

정규화

목적함수

$$L(\beta) = \min_{\beta} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

λ 는 우리가 조절할 수 있는 하이퍼파라미터

LSE와 Generalization Accuracy 사이의

trade-off를 조절하는 역할 수행

3 정규화

정규화

목적함수

$$L(\beta) = \min_{\beta} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Example

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + 5000\beta_3^2 + 5000\beta_4^2$$

위 식과 같이 β_3^2 와 β_4^2 의 계수가 크다면,

expected MSE가 최소가 되게 하기 위해 $\beta_3 \approx 0, \beta_4 \approx 0$ 이 되어야 함

정규화

목적함수

$$L(\beta) = \min_{\beta} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

- ✓ λ가 매우 크다면, $\beta_1 \approx 0, \beta_2 \approx 0, \beta_3 \approx 0, \beta_4 \approx 0 \rightarrow y = \beta_0$ (직선)
- ✓ λ가 매우 작다면, β에 대한 제약이 거의 없는 것

정규화 방법 | ① Ridge (L2 Regularization)

Ridge *L2 Regularization*

SSE를 최소화하면서 회귀계수 β에 제약조건을 거는 방법

제약 조건식이 L2-norm 형태

왜 L2 Regularization인가

제약조건식이 L2-norm 형태 → L2 Regularization

$$L_2 = \sqrt{|v_1|^2 + |v_2|^2 + \cdots + |v_n|^2}$$

선대팀 2주차 클린업 참고

3 정규화

정규화 방법 | ① Ridge (L2 Regularization)

Ridge *L2 Regularization*

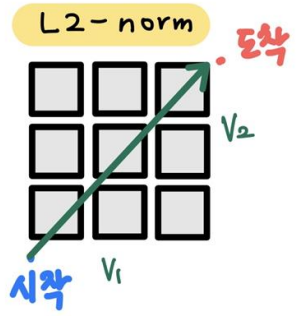
SSE를 최소화하면서 회귀계수 β에 제약조건을 거는 방법

제약 조건식이 L2-norm 형태

왜 L2 Regularization인가

제약조건식이 L2-norm 형태 → L2 Regularization

$$L_2 = \sqrt{|v_1|^2 + |v_2|^2 + \cdots + |v_n|^2}$$



선대팀 2주차 클린업 참고~

정규화 방법 | ① Ridge

목적함수

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_i)^2 \ subject\ to \sum_{j=1}^{p} \beta_j^2 \leq s$$

$$\Leftrightarrow \hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

목적함수를 최소화함으로써 **Ridge Estimator 추정 가능**

이차식 형태이므로 미분을 통해 추정량 계산

정규화 방법 | ① Ridge

목적함수

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_i)^2 \ subject\ to \sum_{j=1}^{p} \beta_j^2 \leq s$$

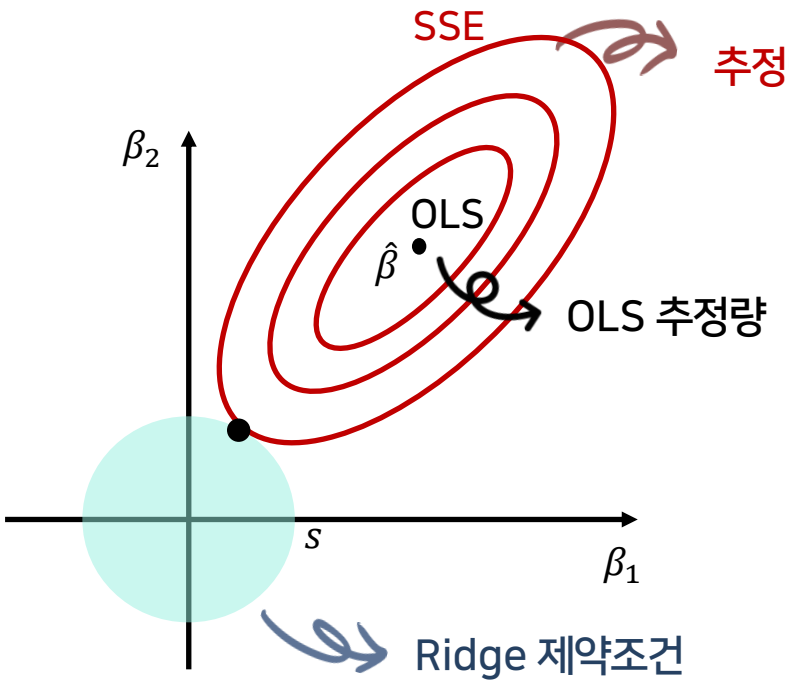
$$\Leftrightarrow \hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

목적함수를 최소화함으로써 Ridge Estimator 추정 가능
이차식 형태이므로 미분을 통해 추정량 계산

단, 설명 변수들은 **표준화된 상태**여야 함!

Ridge | 목적함수에 대한 이해 ①

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_i)^2 \ subject\ to \sum_{j=1}^{p} \beta_j^2 \leq s$$



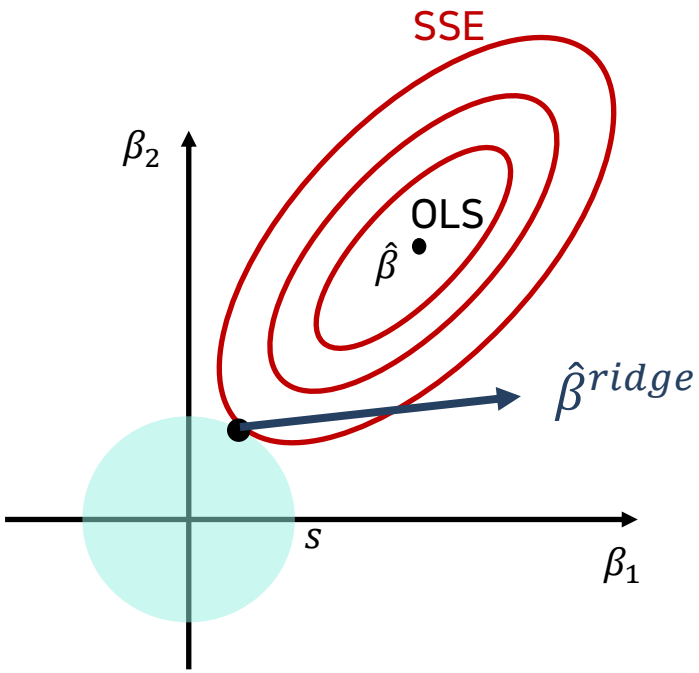
회귀계수의 최소화

① 회귀계수 $\hat{\beta}$는 반드시 원 내부에 존재

② SSE를 최소화해야 함

Ridge | 목적함수에 대한 이해 ①

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_i)^2 \ subject\ to \sum_{j=1}^{p} \beta_j^2 \leq s$$



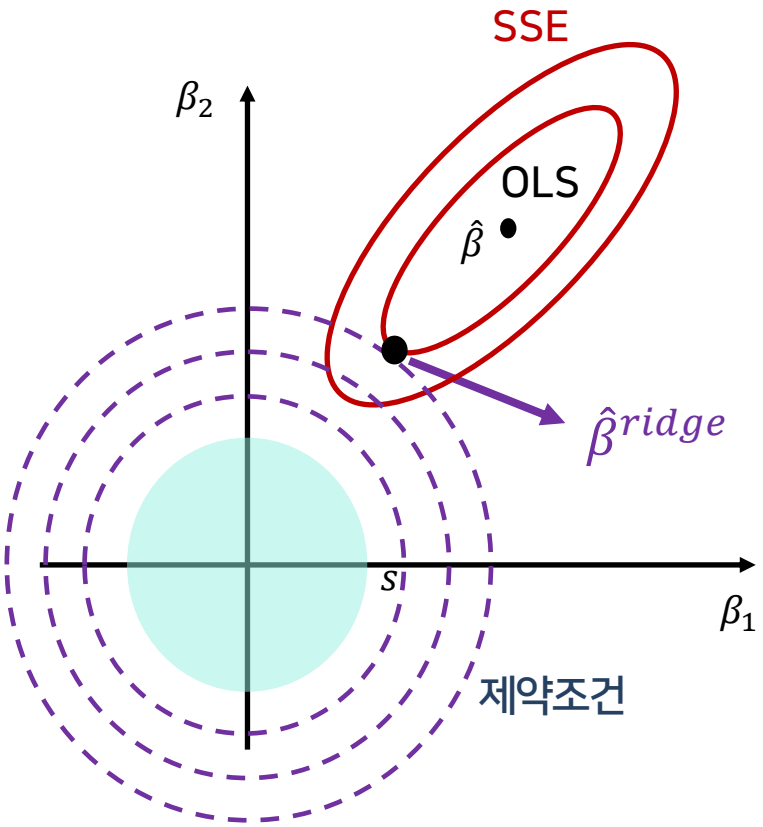
회귀계수의 최소화

타원과 원의 접점이

Ridge Estimator

3 정규화

Ridge | 목적함수에 대한 이해 ①



제약조건이 완화될 경우

s가 증가하면서 원의 넓이 증가

원이 타원을 밀어내며 추정량이 0에서 멀어짐

회귀 계수를 작게 만들 수 없음

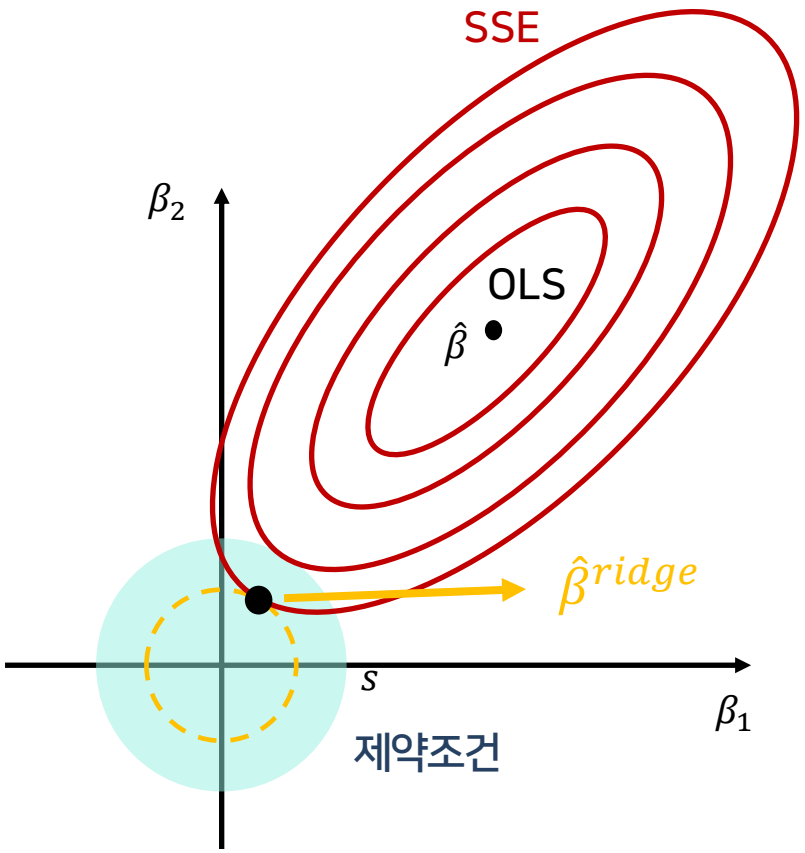
제약조건이 강화될 경우

s가 감소하면서 원의 넓이 감소

추정량이 0으로 수렴함 (0은 될 수 없음)

회귀 계수를 작게 만들 수 있음

Ridge | 목적함수에 대한 이해 ①



제약조건이 완화될 경우

s가 증가하면서 원의 넓이 증가

원이 타원을 밀어내며 추정량이 0에서 멀어짐

회귀 계수를 작게 만들 수 없음

제약조건이 강화될 경우

s가 감소하면서 원의 넓이 감소

추정량이 0으로 수렴함 (0은 될 수 없음)

회귀 계수를 작게 만들 수 있음

Ridge | 목적함수에 대한 이해 ②

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

라그랑주 승수법을 이용해 나타낸 함수식

오차제곱합(SSE) 최소화 & Regularization term을 통해

개별 회귀계수 크기 조정

Ridge | 목적함수에 대한 이해 ②

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

라그랑주 승수법을 이용해 나타낸 함수식

음수가 아닌 **튜닝 파라미터**

최적의 모델을 찾는 과정에서 직접 CV를 통해 조정해주는 모수

제약조건의 크기를 결정 (s와는 반대 관계)

CV: Cross Validation

3 정규화

Ridge | 목적함수에 대한 이해 ②

λ의 값에 따른 회귀계수의 변화

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

λ가 커지는 경우

λ의 영향력이 증가하므로,

전체 식을 최소화하기 위해

$\sum_{j=1}^{p} \beta_j^2$은 작아져야 함

⇒ 개별 회귀 계수들은 감소

λ가 작아지는 경우

λ의 영향력이 감소하므로,

상대적으로 $\sum_{j=1}^{p} \beta_j^2$의 영향력 증가

(s와는 반대 관계)

⇒ 개별 회귀 계수들은 증가

Ridge | 목적함수에 대한 이해 ②

λ의 값에 따른 회귀계수의 변화

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

λ가 커지는 경우

$\lambda \to \infty$

개별 회귀계수의 영향력은
무시될 만큼 작아짐

⇒ 회귀 계수 ≈ 0

λ가 작아지는 경우

$\lambda = 0$

Regularization term이 없어짐

⇒ OLS 추정량과 동일

Ridge | 특징

Scaling

회귀계수는 변수 단위에 큰 영향을 받음

단위의 영향을 제거, 순수 영향력만 사용

주로 standard scaling 사용

계산 비용 절약

Regularization term 덕분에 미분 가능

λ를 바꾸며 미분과 함께 행렬 연산

예측 성능

상관관계가 높은 변수들이

모델에 존재할 경우,

좋은 예측 성능 보임

변수 선택

영향력을 줄일 뿐 변수는 잔존

다중공선성을 일으키는 변수 제거 불가

Ridge를 통한 해석력 증가는 어려움

Ridge | 특징

Scaling

회귀계수는 변수 단위에 큰 영향을 받음

단위의 영향을 제거, 순수 영향력만 사용

주로 standard scaling 사용

계산 비용 절약

Regularization term 덕분에 미분 가능

λ를 바꾸며 미분과 함께 행렬 연산

예측 성능

상관관계가 높은 변수들이

모델에 존재할 경우,

좋은 예측 성능 보임

변수 선택

영향력을 줄일 뿐 변수는 잔존

다중공선성을 일으키는 변수 제거 불가

Ridge를 통한 해석력 증가는 어려움

3 정규화

Ridge | 특징

행렬연산을 통한 closed form solution

$$Q(\beta) = (y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta$$

$$\rightarrow \frac{\partial}{\partial y}Q(\beta) = 2X^Ty + 2(X^TX + \lambda I_p)\beta = 0$$

$$\hat{\beta}^{ridge} = (X^TX + \lambda I_p)^{-1}\ X^Ty \quad \text{vs.} \quad \hat{\beta}^{LS} = (X^TX)^{-1}\ X^Ty$$

I_p는 $p \times p$ 크기의 Identity Matrix이므로

각 λ만큼 더해주었다고 이해할 수 있음!

3 정규화

Ridge | 특징

행렬연산을 통한 closed form solution

Regularization term 덕분에 미분 가능

λ를 바꾸며 미분과 함께 행렬 연산

LS는 BLUE에 의해 unbiased

Ridge는 λ를 더해 주었기에 biased

그러나 variance가 작기 때문에 예측에 더 좋음

Ridge | 특징

Scaling

회귀계수는 변수 단위에 큰 영향을 받음

단위의 영향을 제거, 순수 영향력만 사용

주로 standard scaling 사용

계산 비용 절약

Regularization term 덕분에 미분 가능

λ를 바꾸며 미분과 함께 행렬 연산

예측 성능

상관관계가 높은 변수들이

모델에 존재할 경우,

좋은 예측 성능 보임

변수 선택

영향력을 줄일 뿐 변수는 잔존

다중공선성을 일으키는 변수 제거 불가

Ridge를 통한 해석력 증가는 어려움

Ridge | 특징

Scaling

회귀계수는 변수 단위에 큰 영향을 받음

단위의 영향을 제거, 순수 영향력만 사용

주로 standard scaling 사용

계산 비용 절약

Regularization term 덕분에 미분 가능

λ를 바꾸며 미분과 함께 행렬 연산

예측 성능

상관관계가 높은 변수들이

모델에 존재할 경우,

좋은 예측 성능 보임

변수 선택

영향력을 줄일 뿐 변수는 잔존

다중공선성을 일으키는 변수 제거 불가

Ridge를 통한 해석력 증가는 어려움

3 정규화

Lasso (L1 Regularization)

Lasso *L1 Regularization*

SSE를 최소화하면서 회귀계수 β에 제약조건을 거는 방법

제약 조건식이 L1-norm 형태

왜 L1 Regularization인가

제약조건식이 L1-norm 형태 → L1 Regularization

$$L_1 = |v_1| + |v_2| \cdots + |v_n|$$

선대팀 2주차 클린업 참고~

3 정규화

Lasso (L1 Regularization)

Lasso *L1 Regularization*

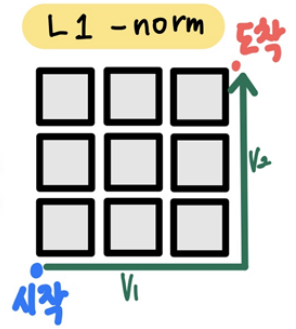
SSE를 최소화하면서 회귀계수 β에 제약조건을 거는 방법

제약 조건식이 L1-norm 형태

왜 L1 Regularization인가

제약조건식이 L1-norm 형태 → L1 Regularization

$$L_1 = |v_1| + |v_2| \cdots + |v_n|$$



선대팀 2주차 클린업 참고~

3 정규화

Lasso (L1 Regularization)

목적함수

$$\hat{\beta}^{Lasso} = argmin\ \Sigma_{\mathrm{i}=1}^{n}\left(y_i - \beta_0 - \Sigma_{j=1}^{p}\beta_j x_i\right)^2 \text{ subject to } \Sigma_{j=1}^{p}|\beta_{\bar{j}}| \le s$$

$$\Leftrightarrow\ \hat{\beta}^{Lasso} = argmin\ \Sigma_{\mathrm{i}=1}^{n}\left(y_i - \beta_0 - \Sigma_{j=1}^{p}\beta_j x_i\right)^2 + \lambda^2\ \Sigma_{j=1}^{p}|\beta_{\bar{j}}|$$

목적함수를 최소화함으로써 Lasso Estimator 추정 가능

단, 설명변수는 **표준화된 형태**여야 함!

3 정규화

Lasso (L1 Regularization)

목적함수

Lasso는 미분이 불가능

$$\hat{\beta}^{Lasso} = argmin \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq s$$

$$\Leftrightarrow \hat{\beta}^{Lasso} = argmin \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 + \lambda^2 \sum_{j=1}^{p} |\beta_j|$$

수치적인 방법을 이용해 최적화 문제 해결

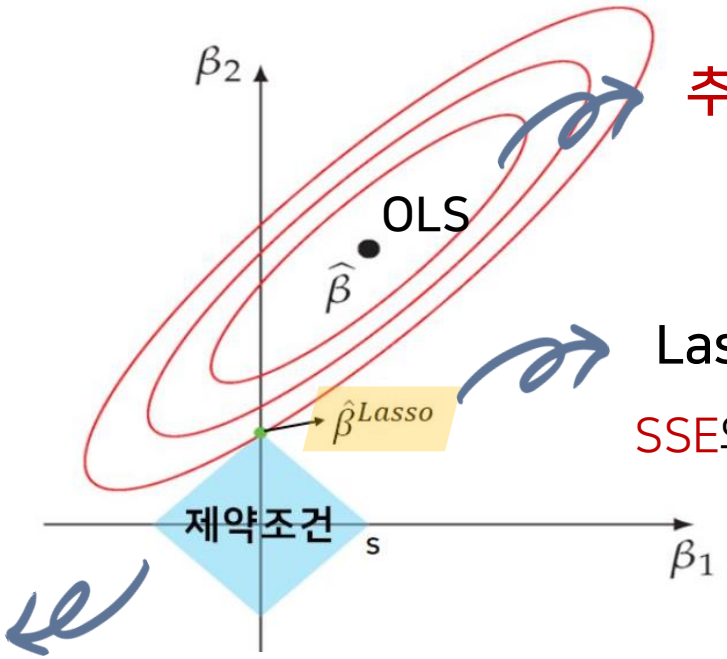
목적함수를 최소화함으로써 Lasso Estimator 추정 가능

단, 설명변수는 표준화된 형태여야 함!

3 정규화

Lasso 목적함수

$$\hat{\beta}^{Lasso} = \underset{\beta}{argmin}\, \Sigma_{i=1}^{n} \left(y_i - \beta_0 - \Sigma_{j=1}^{p} \beta_j x_i\right)^2 \text{ subject to } \Sigma_{j=1}^{p} |\beta_j| \leq s$$



추정량의 SSE

Lasso estimator

SSE와 제약조건의 접점

SSE를 최소화 해야 하기 때문!

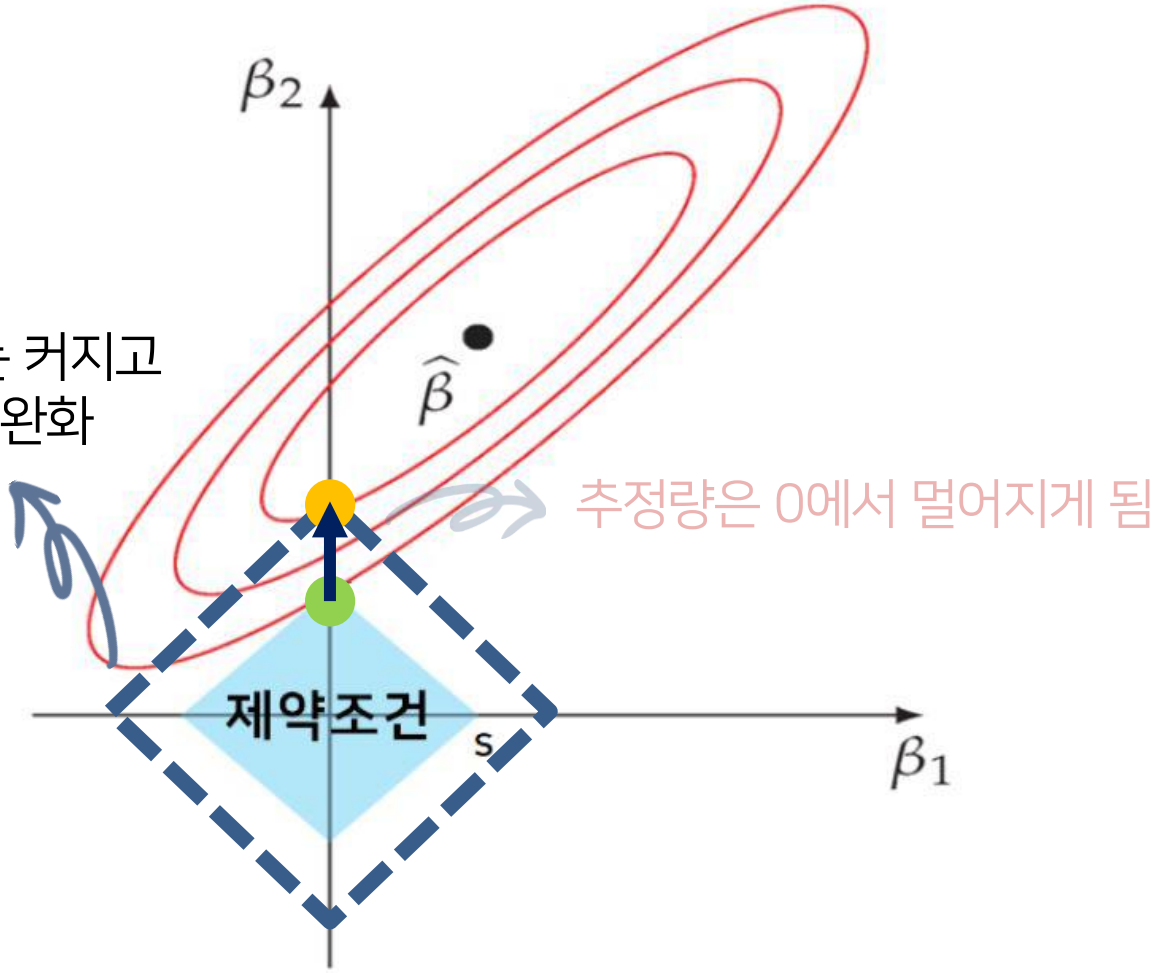
Lasso의 제약조건

3 정규화

Lasso 목적함수

만약 s가 커진다면

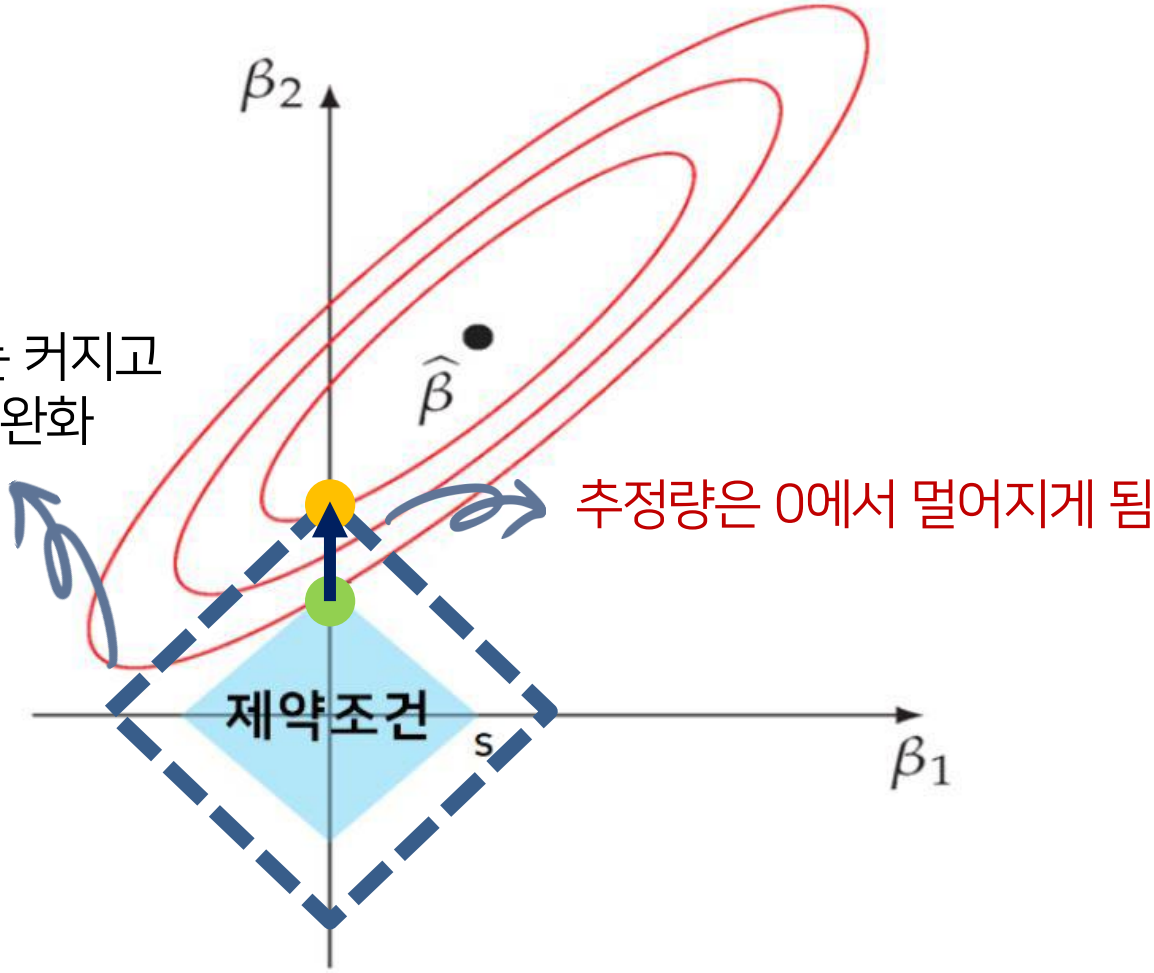
마름모의 넓이는 커지고
제약 조건은 완화



Lasso 목적함수

만약 s가 커진다면

마름모의 넓이는 커지고
제약 조건은 완화

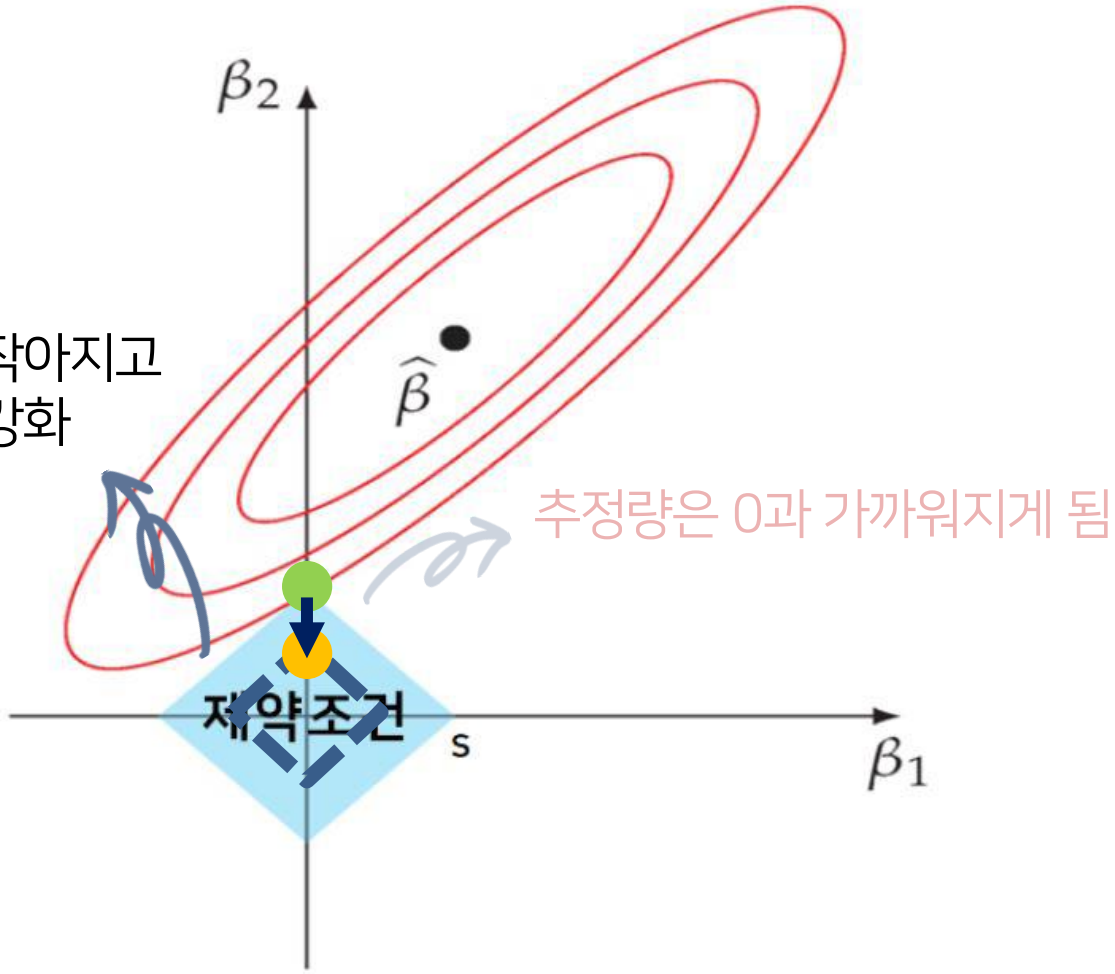


3 정규화

Lasso 목적함수

만약 s가 작아진다면

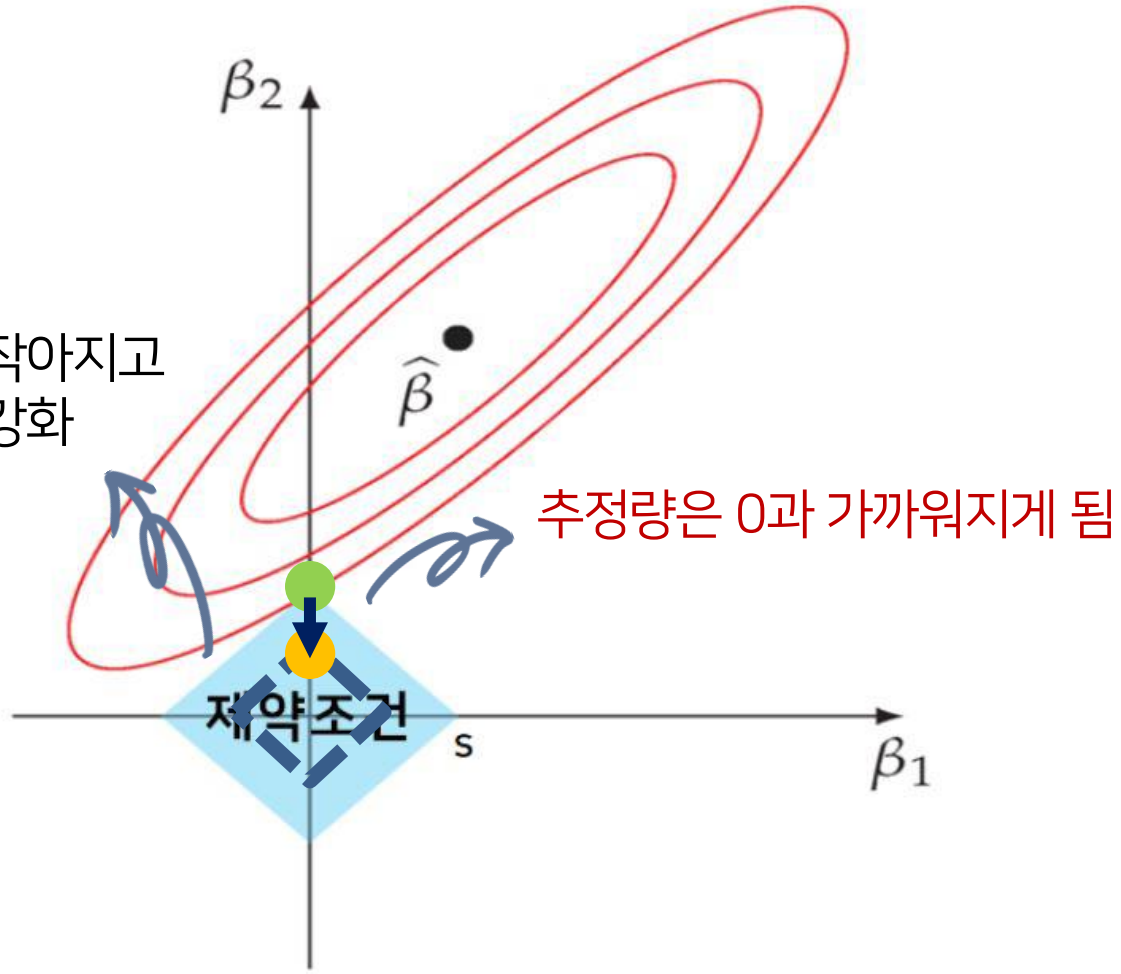
마름모의 넓이는 작아지고
제약 조건은 강화



Lasso 목적함수

만약 s가 작아진다면

마름모의 넓이는 작아지고
제약 조건은 강화

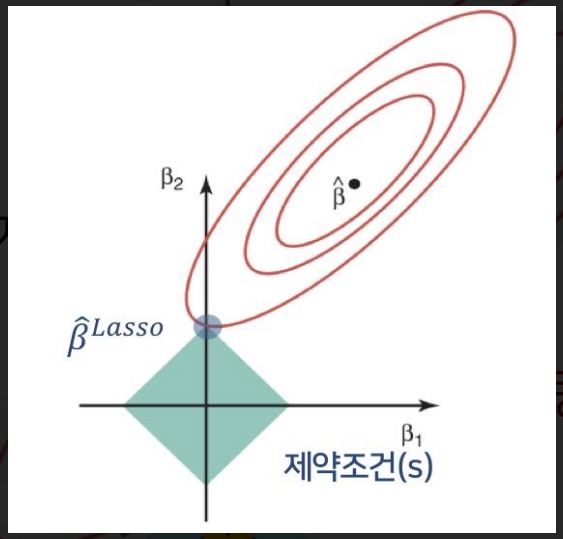


3 정규화

Lasso 목적함수

만약 s가 작아진다면

Ridge의 원리와 매우 비슷!



마름모의 넓이는 작아지
제약 조건은 강화

량은 0과 가까워지게 됨

하지만 Lasso는 일부 회귀계수가 0이 되는 추정량이 도출될 수 있음

$\beta_1 = 0$으로 나온다면 y를 예측하는데 있어 x_1가 유의하지 않다는 것!

따라서 변수 선택 과정에서 활용할 수도 있음

3 정규화

Lasso 목적함수

$$\hat{\beta}^{Lasso} = \underset{\beta}{argmin}\, \Sigma_{i=1}^{n} \left(y_i - \beta_0 - \Sigma_{j=1}^{p} \beta_j x_i\right)^2 \text{ subject to } \Sigma_{j=1}^{p} |\beta_j| \leq s$$

라그랑주 승수법 이용

$$\hat{\beta}^{Lasso} = \underset{\beta}{argmin}\, \Sigma_{i=1}^{n} \left(y_i - \beta_0 - \Sigma_{j=1}^{p} \beta_j x_i\right)^2 + \lambda^2 \Sigma_{j=1}^{p} |\beta_j|$$

오차제곱합 (SSE) term

Regularization term

개별 회귀계수가

너무 많아지는 것을 조정해줌

λ는 CV를 통해 최적값을 찾음

3 정규화

Lasso 목적함수

$$\hat{\beta}^{Lasso} = \underset{\beta}{argmin} \Sigma_{i=1}^{n} \left(y_i - \beta_0 - \Sigma_{j=1}^{p} \beta_j x_i\right)^2 \text{ subject to } \Sigma_{j=1}^{p} |\beta_j| \leq s$$

λ의 값에 따른 회귀계수의 변화

λ가 커지는 경우

λ의 영향력이 증가하므로,
전체 식을 최소화하기 위해
$\Sigma_{j=1}^{p} \beta_j^2$은 작아져야 함
⇒ 개별 회귀 계수들은 감소

λ가 작아지는 경우

λ의 영향력이 감소하므로,
상대적으로 $\Sigma_{j=1}^{p} \beta_j^2$의 영향력 증가
너무 많아지는 것을 조정해줌
⇒ 개별 회귀 계수들은 증가

3 정규화

Lasso 목적함수

λ의 값에 따른 회귀계수의 변화

$$\hat{\beta}^{Lasso} = \underset{\beta}{argmin} \Sigma_{i=1}^{n} \left(y_i - \beta_0 - \Sigma_{j=1}^{p} \beta_j x_i\right)^2 \text{ subject to } \Sigma_{j=1}^{p} |\beta_j| \leq s$$

라그랑주 승수법 이용

λ가 커지는 경우

$\lambda \to \infty$

오차제곱합 (SSE) term

개별 회귀계수의 영향력은

무시될 만큼 작아짐

⇒ 회귀 계수 ≈ 0

λ가 작아지는 경우

$\lambda = 0$

Regularization term

Regularization term이 없어짐

개별 회귀계수가

너무 많아지는 것을 조정해줌

⇒ OLS 추정량과 동일

λ는 CV를 통해 최적값을 찾음

3 정규화

Lasso 목적함수

큰 λ값	작은 λ값
적은 변수	많은 변수
간단한 모델	복잡한 모델
해석 쉬움	해석 어려움
높은 학습 오차 (underfitting 위험 증가)	낮은 학습 오차 (overfitting 위험 증가)

Lasso 특징

Scaling

개별 변수들에 대한 scaling 필요

변수의 단위에 의한 영향력 제거

주로 standard scaling 사용

변수 선택

0이 되는 회귀 계수가 존재해 변수 선택 가능

변수 선택으로 변수들의 해석 가능성 증가

예측 성능

변수들 간 상관관계가 큰 경우,

유의미한 변수들을 0으로 만들 수 있음

Ridge 보다 상대적으로 예측 성능 저하

Closed Form Solution

미분이 불가능한 점이 존재하여

Closed form solution을 구할 수 없음

➡ **수치 최적화 방법** 사용

ex) Quadratic programming technique,
LARS · Coordinate descent algorithm

3 정규화

Lasso 특징

Scaling

개별 변수들에 대한 scaling 필요

변수의 단위에 의한 영향력 제거

주로 standard scaling 사용

변수 선택

0이 되는 회귀 계수가 존재해 변수 선택 가능

변수 선택으로 변수들의 해석 가능성 증가

변수간 상관관계가 높으면 변수 선택 성능이 떨어짐!

0이 되는 계수의 존재로 인해 **sparsity**를 지님

Lasso 특징

Scaling

개별 변수들에 대한 scaling 필요

변수의 단위에 의한 영향력 제거

주로 standard scaling 사용

변수 선택

0이 되는 회귀 계수가 존재해 변수 선택 가능

변수 선택으로 변수들의 해석 가능성 증가

예측 성능

변수들 간 상관관계가 큰 경우,

유의미한 변수들을 0으로 만들 수 있음

Ridge 보다 상대적으로 예측 성능 저하

Closed Form Solution

미분이 불가능한 점이 존재하여

Closed form solution을 구할 수 없음

➡ 수치 최적화 방법 사용

ex) Quadratic programming technique,
LARS · Coordinate descent algorithm

3 정규화

Lasso 특징

Scaling

개별 변수들에 대한 scaling 필요

변수의 단위에 의한 영향력 제거

주로 standard scaling 사용

변수 선택

0이 되는 회귀 계수가 존재해 변수 선택 가능

변수 선택으로 변수들의 해석 가능성 증가

예측 성능

변수들 간 상관관계가 큰 경우,

유의미한 변수들을 0으로 만들 수 있음

Ridge 보다 상대적으로 예측 성능 저하

Closed Form Solution

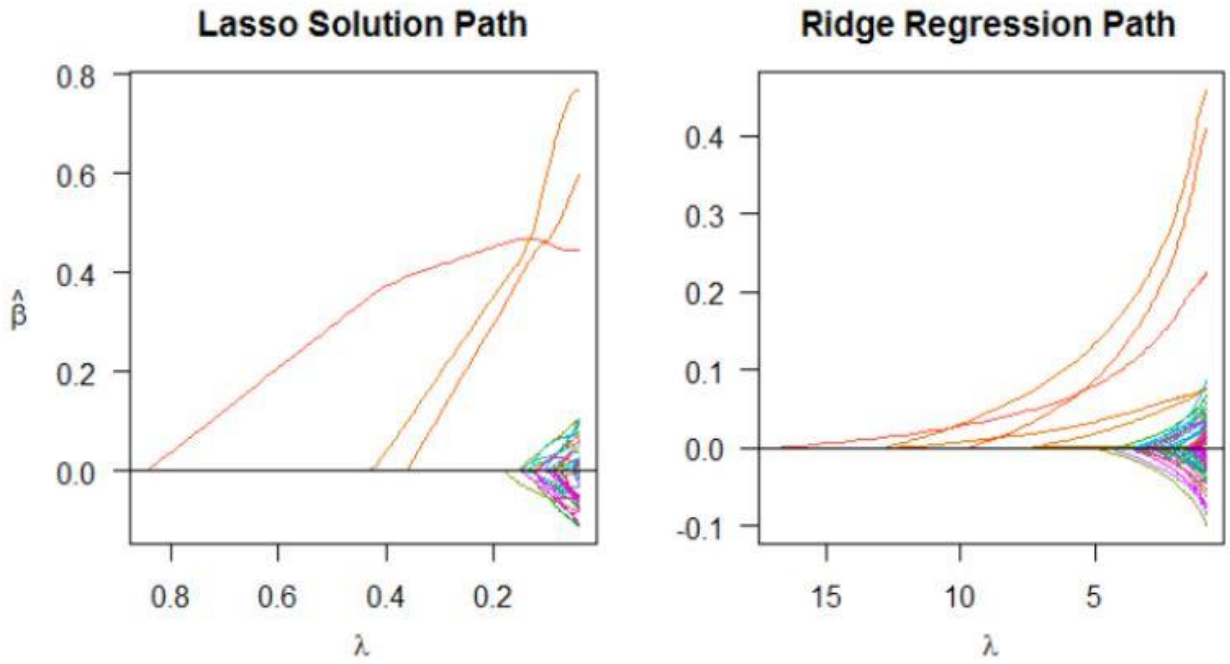
미분이 불가능한 점이 존재하여

Closed form solution을 구할 수 없음

수치 최적화 방법 사용

ex) Quadratic programming technique,
LARS · Coordinate descent algorithm

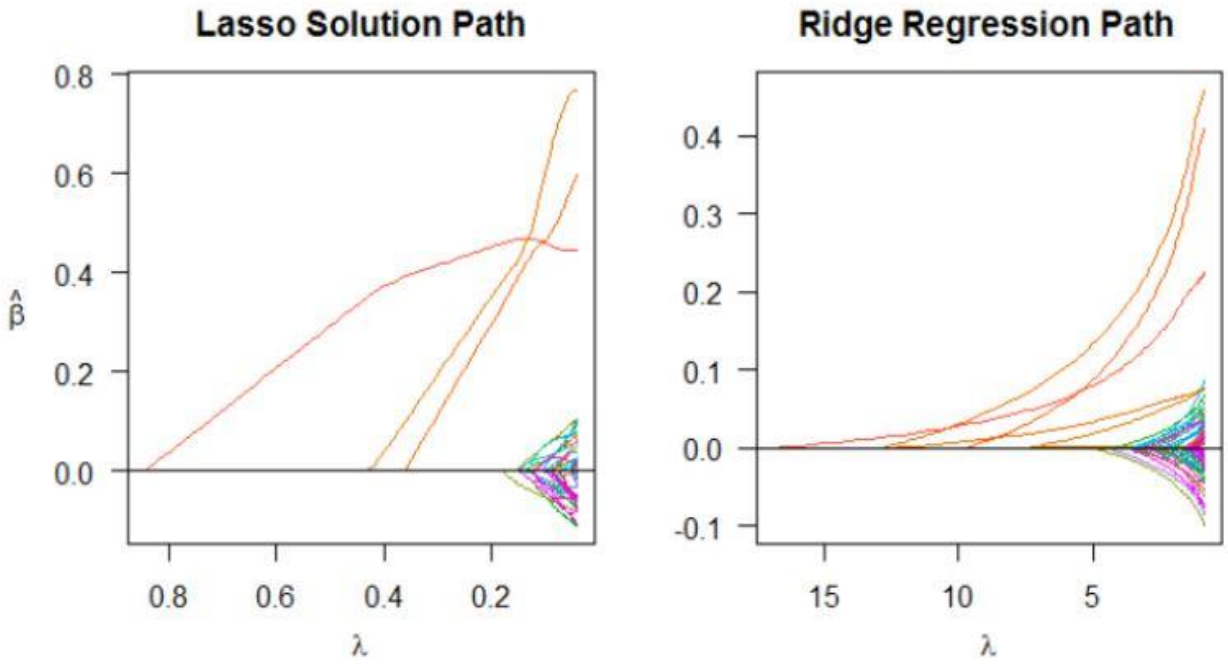
Lasso와 Ridge 비교



공통점

Lasso와 Ridge 모두 λ가 커짐에 따라 모든 계수의 크기가 감소

Lasso와 Ridge 비교



차이점

- ✓ Lasso는 예측에 중요하지 않은 변수가 더 빠르게 감소
- ✓ Lasso는 λ가 커짐에 따라 예측에 중요하지 않은 변수가 0이 됨

3 정규화

Lasso와 Ridge 비교

Ridge regression	Lasso regression
변수 선택 불가능	변수 선택 가능
Closed form solution 존재 (미분을 통해)	Closed form solution 존재 X (수치최적화 방법 사용)
상관관계가 높은 상황에서 좋은 예측	변수 간 상관관계가 높은 상황에서 Ridge에 비해 상대적으로 예측 성능 저하
제약 범위 = 원	제약 범위 = 사각형

3 정규화

Elastic-Net

Elastic-Net *Elastic-Net*

Ridge와 Lasso의 **regularization term을 혼합**한 형태

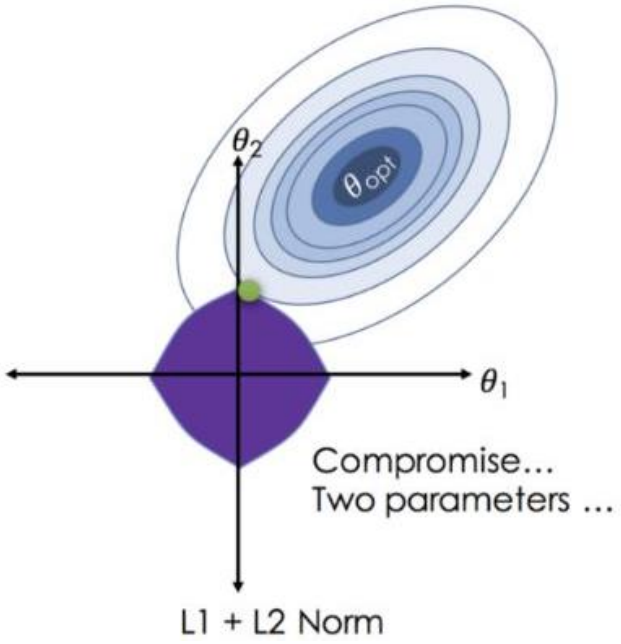
변수 간 상관관계가 존재할 때

Lasso의 성능이 떨어지는 한계를 보완하기 위한 방법

Elastic-Net

Elastic-Net *Elastic-Net*

Ridge와 Lasso의 regularization term을 혼합한 형태



Grouping Effect

상관성이 있는 변수를

모두 선택하거나 제거해 성능 보완

3 정규화

Elastic-Net

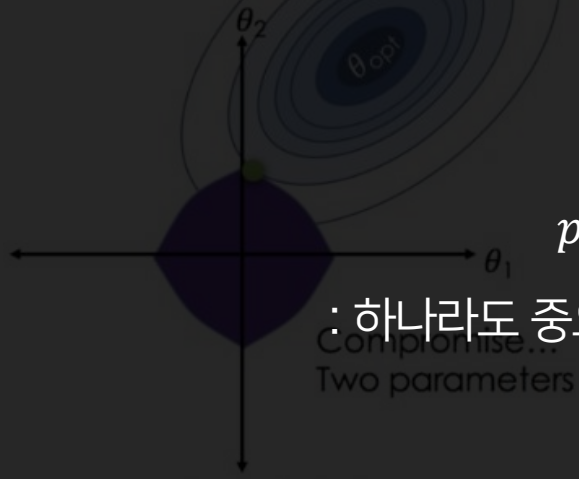
Elastic-Net

Grouping Effect 추가 설명

Ridge와 Lasso의 장점을 혼합한 형태

$$\left|\hat{\beta}_i^{\text{enet}} - \hat{\beta}_j^{\text{enet}}\right| \leqq \frac{\Sigma_{i=1}^{n}|y_i|}{\lambda_2}\sqrt{2(1-p_{ij})}$$

P_{ij}는 x_i와 x_j의 상관계수를 의미



Grouping Effect

상관성이 있는 변수를 모두 선택하거나 제거해 성능 보완

$p_{ij} = 1 \rightarrow \left|\hat{\beta}_i^{\text{enet}} - \hat{\beta}_j^{\text{enet}}\right|$

: 하나라도 중요하다면 둘 다 똑같이 중요하다는 의미

$\Rightarrow p_{ij}$ 가 증가하거나 λ_2가 증가한다면 $\left|\hat{\beta}_i^{\text{enet}} - \hat{\beta}_j^{\text{enet}}\right|$ 는 감소

3 정규화

Elastic-Net 목적함수

$$\hat{\beta}^{elastic} = \underset{\beta}{argmin}\ \Sigma_{i=1}^{n}\left(y_i - \beta_0 - \Sigma_{j=1}^{p}\beta_j x_i\right)^2 \ \text{s.t.}\ t_1\ \Sigma_{j=1}^{p}|\beta_j| + t_1\ \Sigma_{j=1}^{p}\beta_j^2 \le s$$

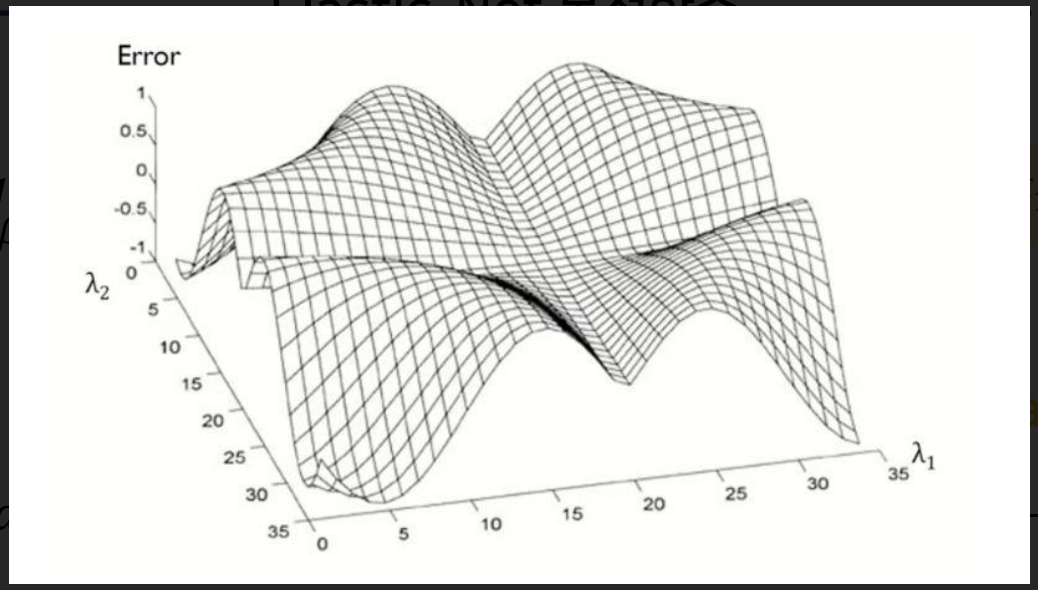
Ridge의 L2 term

Lasso의 L1 term

$$\Leftrightarrow \hat{\beta}^{elastic} = \underset{\beta}{argmin}\ \Sigma_{i=1}^{n}\left(y_i - \beta_0 - \Sigma_{j=1}^{p}\beta_j x_i\right)^2 + \lambda_1\ \Sigma_{j=1}^{p}|\beta_j| + \lambda_2\ \Sigma_{j=1}^{p}\beta_j^2$$

3 정규화

Elastic-Net 목적함수



$\hat{\beta}^{elastic} = arg$ $\quad \Sigma_{j=1}^{p}\beta_j^2 \le s$

asso의 L1 term

$\Leftrightarrow \hat{\beta}^{elastic} =$ $\quad \lambda_2 \Sigma_{j=1}^{p}\beta_j^2$

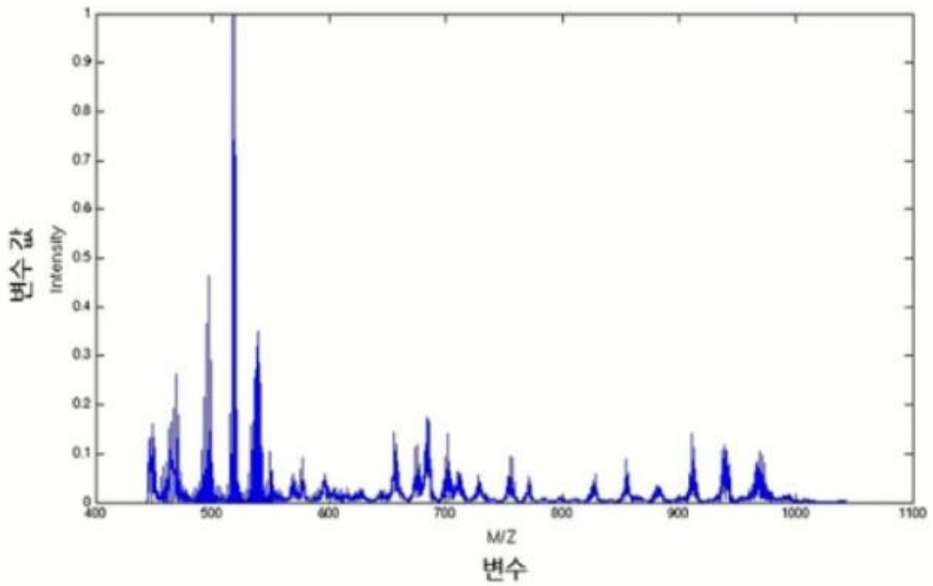
👉 Grid search 방법을 사용해

λ_1와 λ_2의 범위를 정해 error가 최소화되는 조합을 찾아감

Fused Lasso

Fused Lasso *Fused Lasso*

변수들 사이의 **물리적 거리**가 존재한다는 **사전 지식**을 **활용**한 모델



인접한 변수들을 동시에 고려하기 위해 마련

3 정규화

Fused Lasso 목적함수

$$\hat{\beta}^{FL} = \underset{\beta}{argmin}\left(\Sigma_{i=1}^{n}\left(y_i - \beta_0 - \Sigma_{j=1}^{p}\beta_j x_i\right)^2 + \lambda_1 \Sigma_{j=1}^{p}|\beta_j| + \lambda_2 \Sigma_{j=1}^{p}|B_j - \beta_{j-1}|\right)$$

Lasso의 L1 term

새로 추가된 term

상관관계와 관계 없이

물리적으로 인접한 변수들의 회귀계수를

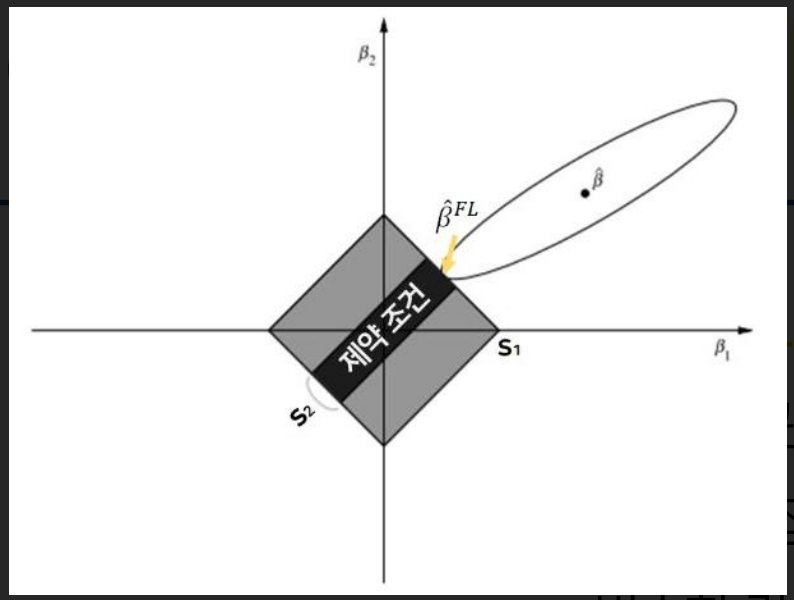
비슷한 값으로 추정하게 만듦

양 옆에 있는 변수들의
회귀계수 값을 최소화 (smoothness)

3 정규화

Fused Lasso 목적함수

$$\hat{\beta}^{FL} = \underset{\beta}{argmin}\left(\Sigma_{i=1}^{n} \quad \lambda_2 \quad \Sigma_{j=1}^{p} |B_j - \beta_{j-1}| \right)$$



추가된 term

관계와 관계 없이

접한 변수들의 회귀계수를

비슷한 값으로 추정하게 만듦

양 옆에 있는 변수들의

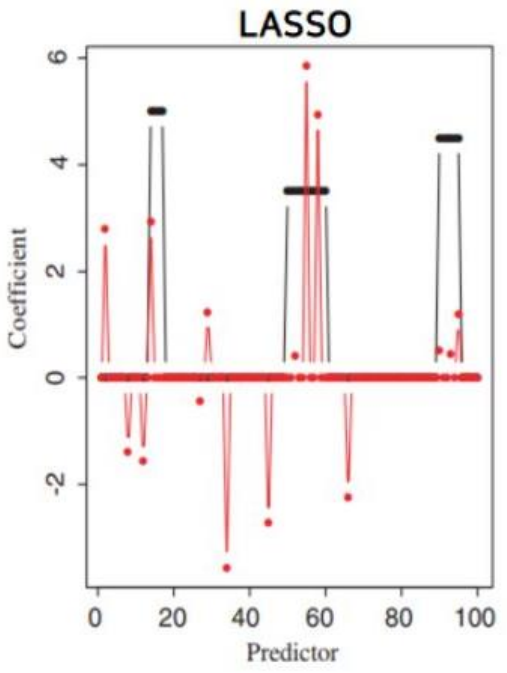
회귀계수 값을 최소화 (smoothness)

변수들의 차이에 대한 제약으로,

인접한 변수의 값을 비슷하게 추정하도록 엄격한 제약 조건을 설정

3 정규화

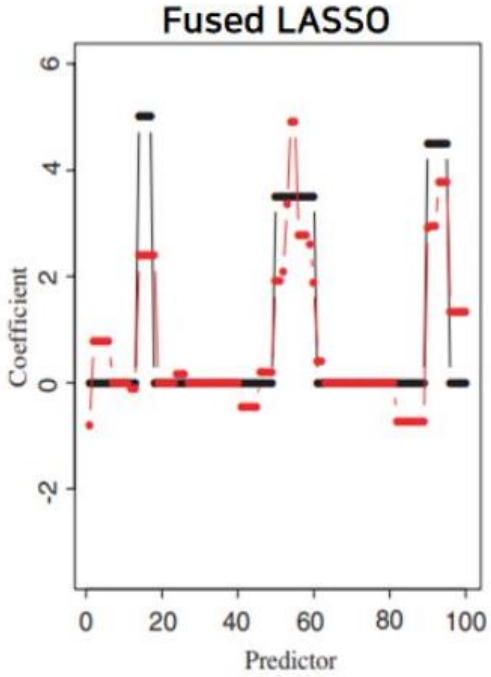
Lasso와 Fused Lasso 비교



Lasso

인접한 실제 계수들을

제대로 추정하지 못함



Fused Lasso

서로 인접한 변수들의 계수가

비슷하게 추정됨

감사합니다