회귀분석팀

6팀

김민우

채희지

김다민

성준혁

천예원

INDEX

1

다중공선성

1 다중공선성

다중공선성이란?

다중공선성

모델에서 설명변수 X_j들 사이에 서로 **선형적인 상관관계**가 존재하는 것

회귀분석의 기본 가정

모델의 선형성	*오차의 정규성*	*오차의 등분산성*	*오차의 독립성*

1 다중공선성

다중공선성이란?

다중공선성

모델에서 설명변수 X_j들 사이에 서로 **선형적인 상관관계**가 존재하는 것

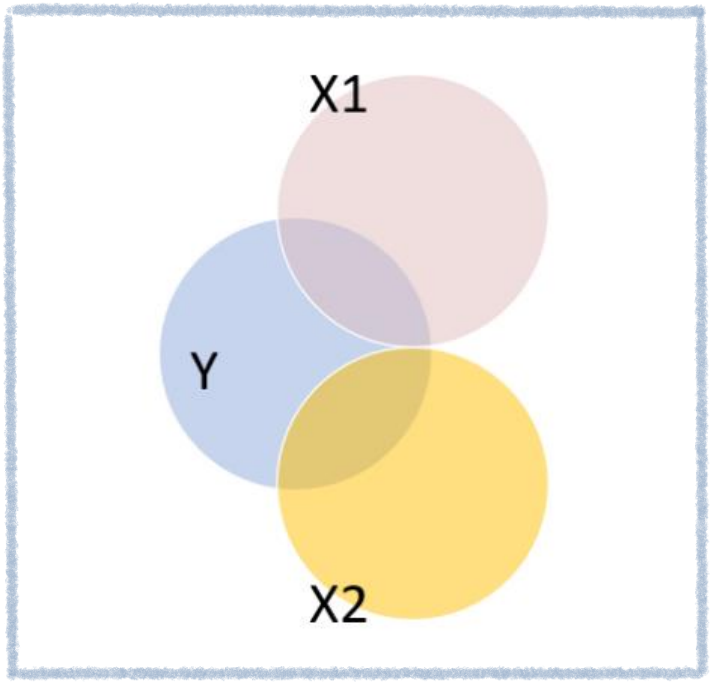
회귀분석의 기본 가정

모델의 선형성 | 오차의 [illegible] | ***설명변수의 독립성*** | 오차의 독립성

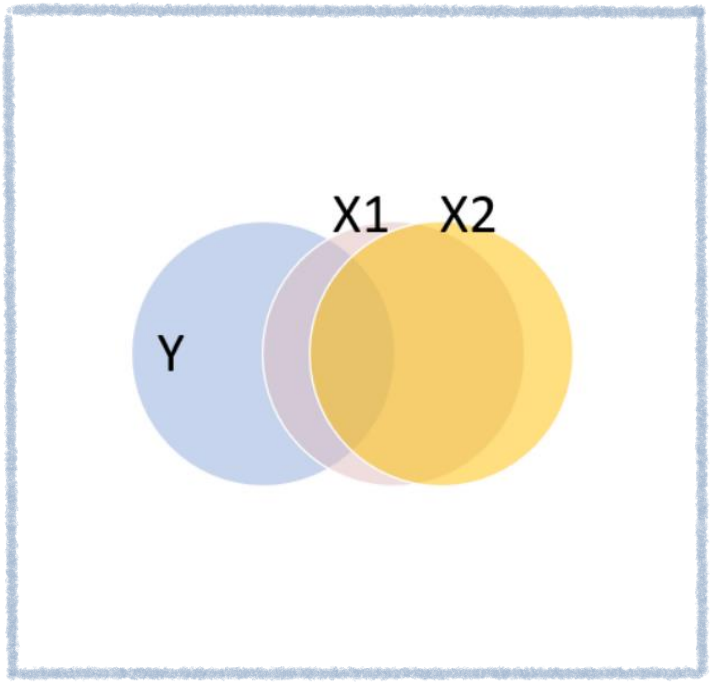
위 가정이 **위배**된 경우가 다중공선성의 경우!

1 다중공선성

다중공선성이란?



▲ 다중공선성이 **없는** 경우



▲ 다중공선성이 **있는** 경우

1 다중공선성

다중공선성이란?

예시

Y : 학점, X_1 : 결석 수, X_2 : 출석률, X_3 : 강의 수

X_2 변수는 $X_2 = \left(1 - \frac{X_1}{X_3}\right)$ 와 같은 식으로 X_1, X_3에 의해 완전히 설명됨

X_2의 정보는 완전히 **필요하지 않은** 정보

1 다중공선성

문제 | ① 추정량의 문제

1) 모수의 추정 자체를 어렵게 만든다

최소제곱법(OLS)을 통한 LSE로 적합된 모형

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

역행렬이 존재하려면?

$X'X$가 full rank → X가 full-column rank → X의 p개 변수가 **선형독립**

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

1 다중공선성

문제 | ① 추정량의 문제

1) 모수의 추정 자체를 어렵게 만든다

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

최소제곱법(OLS)을 사용할 수 없게 됨 → 모수의 추정 자체가 어려워짐

Complete Multicollinearity

(완전한 선형종속)

현실에서는 근사적으로 선형종속을 이루는 경우가 많음!

문제 | ① 추정량의 문제

2) 추정량을 불안정하게 만든다

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

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

이때, 다중공선성이 존재한다면 $\det(X'X) \approx 0$

⋮

추정량의 분산 또한 매우 커져 **계수의 추정이 불안정**해지는 문제 발생

▶ Prediction Accuracy도 심각하게 감소

1 다중공선성

문제 | ② 모델의 문제

1) 모델의 검정 결과를 신뢰할 수 없다

회귀 모델 전체에 대한 검정인 F-test 통과, 적합성 검정을 위한 R^2 값도 괜찮은 수준

⋮

유의한 개별 계수가 **하나도 존재하지 않는** 상황 발생

회귀계수들의 분산이 커짐 > t 검정통계량 감소 > 귀무가설 기각 못함

개별 변수의 유의성 검정에서!

문제 | ② 모델의 문제

2) 모델 해석에 영향을 준다

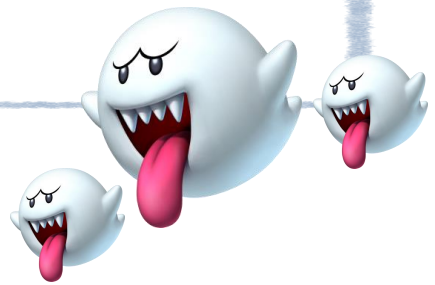
개별 베타계수 β_j 의 해석

β_j

다중선형회귀모델에서,

변수 x_j를 제외한 **나머지 변수가 고정**되어 있을 때,

x_j가 한 단계 증가했을 때의 증가량



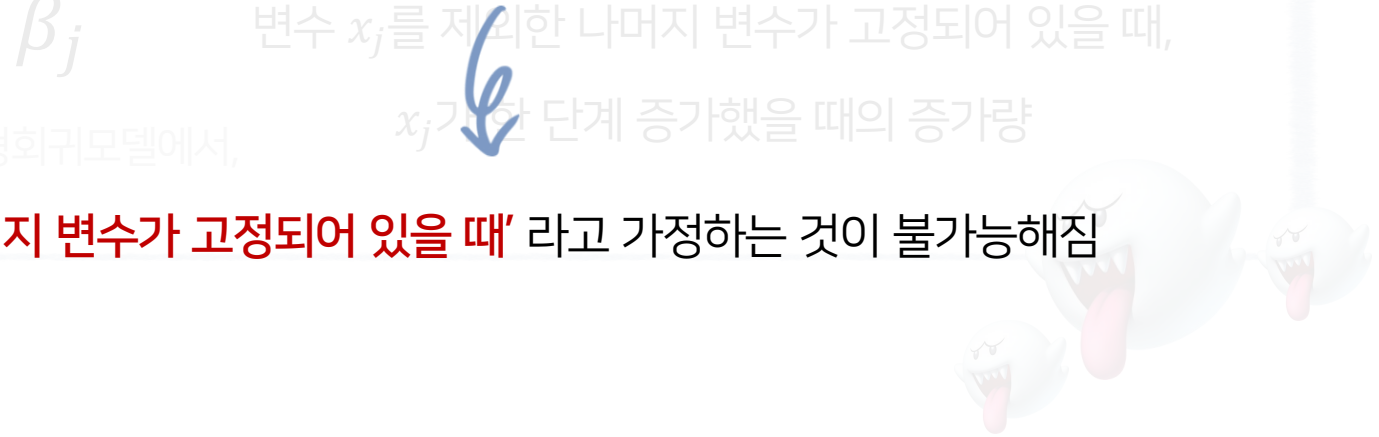
1 다중공선성

문제 | ② 모델의 문제

2) 모델 해석에 영향을 준다

다중공선성이 있는 경우 x_j가 변할 때

선형종속 관계에 있는 다른 변수들도 변할 수 있음



‘나머지 변수가 고정되어 있을 때’ 라고 가정하는 것이 불가능해짐

1 다중공선성

문제 | ② 모델의 문제

모델의 문제

다중공선성 문제가 예측 성능에 영향을 주지 않는 경우

2) 모델 해석에 영향을 준다

▶ 학습 데이터와 검증 데이터에서 변수들의 공분산이 일치하거나 비슷할 경우

▶ 랜덤포레스트, 부스팅 모델 등의 경우

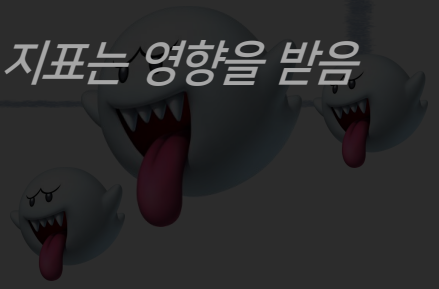
개별 배타계수 β_j 의 해석

β_j 변수 x_j를 제외한 나머지 변수가 고정되어 있을 때,

다중선형회귀모델에서,

Black-Box 모델들은 애초에 Robust한 예측을 목적으로 만들어졌기 때문!

랜덤포레스트의 Feature Importance 지표는 영향을 받음



1 다중공선성

진단 | ① 직관적인 판단

F-test는 유의했지만, 개별 회귀계수들에 대한 T-test가
귀무가설을 대부분 기각하지 못할 때

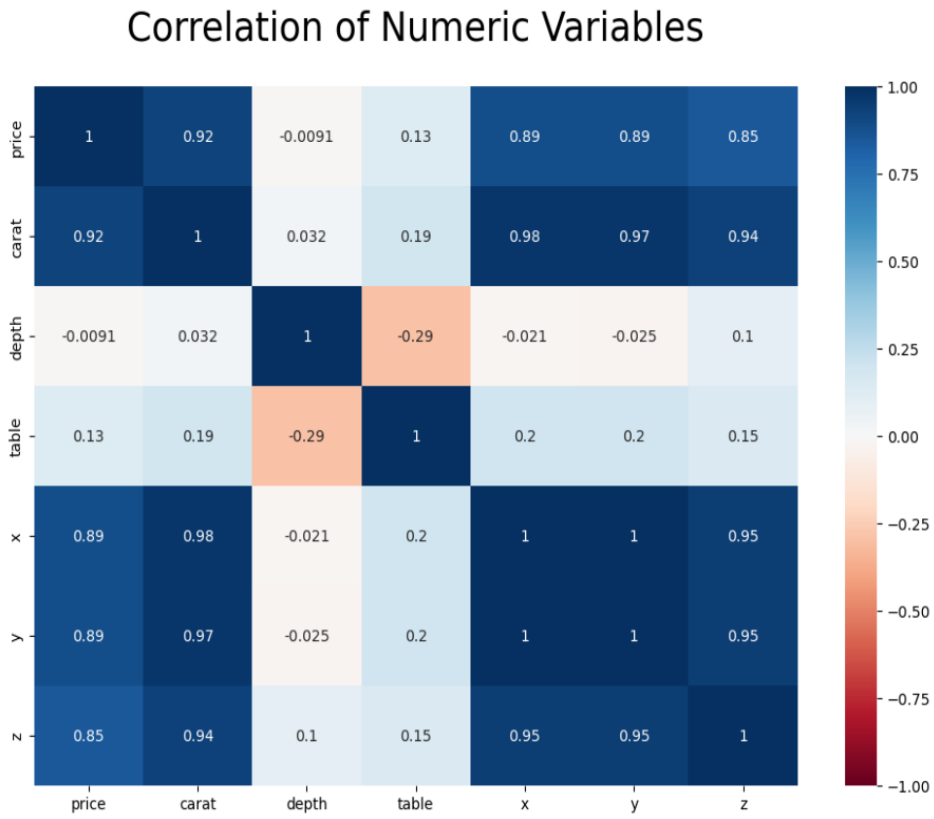
상식적으로 유의한 회귀계수가, 유의하지 않다고 나올 때

이미 비슷한 설명변수가 모델에 포함되어 있을지도…!

추정된 회귀계수의 부호가, 상식과 다르게 나올 때

1 다중공선성

진단 | ② 상관계수 Plot



- ✓ 변수들 사이의 선형관계 파악 가능
- ✓ 상관계수 **절댓값이 0.7 이상**일 때 다중공선성 의심!

패키지 단골 손님 ^_^

진단 | ③ VIF

VIF (분산팽창인자) *Variance Inflation Factor*

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

R_j^2

선형회귀식 $x_j = \gamma_1 x_1 + \cdots + \gamma_{j-1} x_{j-1} + \gamma_{j+1} x_{j+1} + \cdots + \gamma_p x_p$ 을

적합했을 때의 결정계수로, **독립변수들 간 관계**를 설명 가능

R_j^2 이 크다 → x_j 가 **나머지 변수들의 선형결합**으로 충분히 표현될 수 있다 → **다중공선성** 존재!

VIF도 커짐

진단 | ③ VIF

VIF (분산팽창인자) *Variance Inflation Factor*

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

✔ 일반적으로 VIF가 10 이상으로 나오면, 심각한 다중공선성 의미 !

✔ 다중공선성이 전혀 없다면, VIF = 1 !

2 다중공선성

해결 방법

선대팀 클린업 참고~

변수선택법 (Variable Selection)	**차원축소** (Dimension Reduction)
정규화 (Regularization)	**필터링 방법** (Filtering Method)

- ▶ 차원축소 방법에는 PCA, PLS, 신경망 모형을 사용한 AE, 요인분석 등이 있음
- ▶ 필터링 방법은 모델링 이전에 변수 자체의 통계적 특징만으로 변수를 선택

2 다중공선성

해결 방법

변수선택법
(Variable Selection)

(D ction)

정규화
(Regularization)

(Fr d)

▶ 차원축소 방법에는 PCA, PLS, 신경망 모형을 사용한 A 분석 등이 있음

▶ 필터링 방법은 모델링 이전에 변수 자체의 통계적 특징 변수를 선택

2

변수선택법

2 변수선택법

변수선택법이란?

변수 선택법

분석을 위해 고려할 많은 변수들 중 **적절한 변수의 조합**을 찾아내는 방법

우리에게 주어진 후보 변수들(Candidate Regressor) 중에서,

일부분만 중요하거나 예측에 유의미할 수 있음

높은 상관관계를 가지는 변수를 제거하여 **다중공선성** 해결!

2 변수선택법

변수선택법이란?

변수선택법은 다중공선성이 발견되지 않더라도 사용 가능!

분석을 위해 고려할 많은 변수들 중 **적절한 변수의 조합**을 찾아내는 방법

▶ 변수 선택을 통해 모델에 대한 **해석력 증가**

▶ 최종 모델에 대한 **확신 획득**

우리에게 주어진 후보 변수 셋(Candidate Regressor) 중에서,

▶ 최대한 적은 변수를 사용해 **모형의 분산 감소**

즉, 변수 선택과 제거에 논리성과 정당성을 부여하는 방법

변수선택법으로 다중공선성을 완벽히 제거하지는 못할 수 있다는 점도 명심!

2 변수선택법

무슨 기준으로

변수를 선택하는데?

변수 선택 지표 | ① Partial F-test

Partial F-test

일부 회귀계수에 대한 유의성 검정

$$model\ A\ :\ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \text{ (Reduced Model)}$$

$$model\ B\ :\ y = \beta_0' + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \text{ (Full Model)}$$

유의하지 않은 변수들을 없애는 방식으로

변수 선택을 진행할 수 있음!

변수 선택 지표 | ① Partial F-test

Partial F-test

일부 회귀계수에 대한 유의성 검정

하지만, **내포 관계에 있지 않은 모델들을 비교**해야 하는 경우도 존재!

$model\ A: \mathrm{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ vs. $model\ B: \mathrm{y} = \beta_0 + \beta_3 x_3 + \beta_4 x_4$

▶ Partial F-test **사용 불가**

일반적인 상황에서도 (내포관계와 무관하게)

모델 간의 비교를 가능하게 해주는 지표가 필요

2 변수선택법

변수 선택 지표 | ① Partial F-test

변수선택법의 핵심

적은 변수로 데이터를 가장 잘 설명하는 모델을 찾는 것!

▶ **모델의 설명력**

▶ **변수의 개수**

둘 다 고려하는 지표를 살펴보자!

일반적인 상황에서도 (내포관계와 무관하게)

모델 간의 비교를 가능하게 해주는 지표가 필요

변수 선택 지표 | ② 수정결정계수

수정결정계수 (R^2_{adj})

설명력을 담당하는 **결정계수, 변수 개수에 대한 페널티** 복합적으로 고려 가능

R^2_{adj} 계산식

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

SSE : 오차의 제곱합 / SST : 전체 제곱합 / p : 변수의 개수

변수 선택 지표 | ③ AIC

AIC *(Akaike Information Criterion)*

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

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

⋮

$Likelihood$ 가 커지면 AIC는 작아짐

▶ AIC가 **낮을수록 더 좋은 모형**으로 해석!

변수 선택 지표 | ③ AIC

AIC *(Akaike Information Criterion)*

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

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

p : 모델의 모수 개수

▶ **변수의 개수**에 따른 **페널티** 부과!

변수 선택 지표 | ④ BIC

BIC *(Bayesian Information Criterion)*

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

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

p : 모델의 모수 개수 / n : 데이터의 개수

AIC와 마찬가지로,

낮을수록 더 좋은 모형으로 해석!

변수 선택 지표 | ④ BIC

BIC *(Bayesian Information Criterion)*

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

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

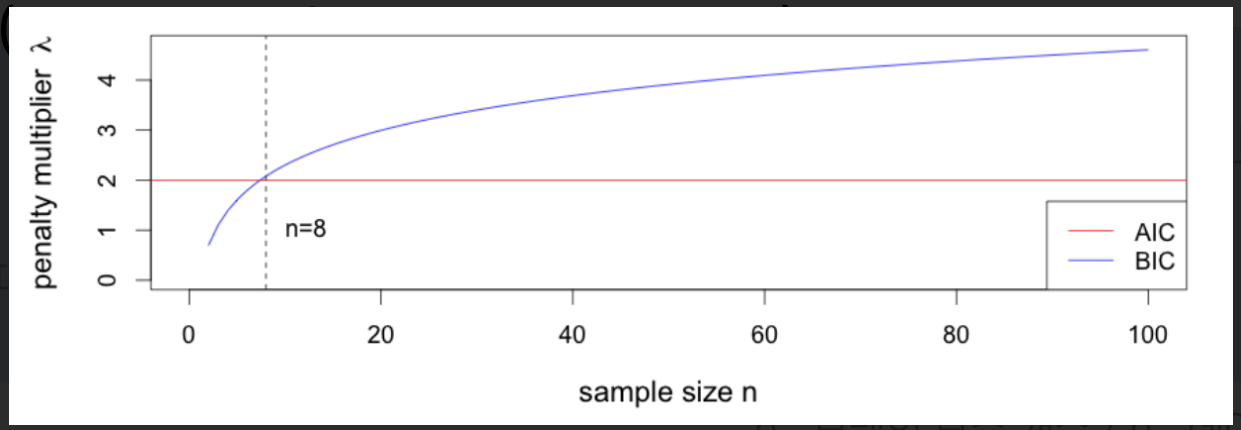
p : 모델의 모수 개수 / n : 데이터의 개수

변수의 개수에 데이터의 개수를 곱해
AIC보다 **더 큰 페널티를 부과**한다는
차이가 있음!

2 변수선택법

변수 선택 지표 | ④ BIC

BIC (



n > 8이라면, **BIC**가 AIC보다 **더 많은 페널티**를 부여

변수의 개수에 데이터의 개수를 곱해

AIC보다 더 큰 페널티를 부과한다.

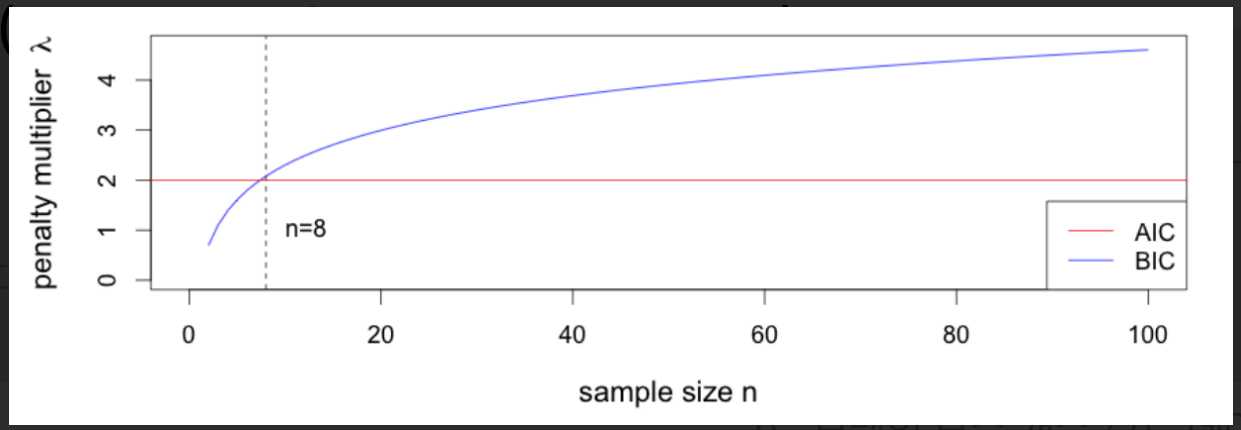
변수의 개수가 적은 것이 우선순위라면 **BIC**를 참고하는 것이 좋음!

차이가 있음!

2 변수선택법

변수 선택 지표 | ④ BIC

BIC (



단, **고차원 데이터**에서는 정확성이 떨어질 수 있고

AIC와 BIC 모두에서 문제가 발생할 수 있어

두 지표를 **종합적으로 고려**해서 모형을 선택해야 함!

변수 선택 방법

변수선택법은 모두 경험적인 방법

직접 모든 경우를 계산해서

제일 좋은 회귀식을 찾는 방법

⋮

계산량이 많다 …

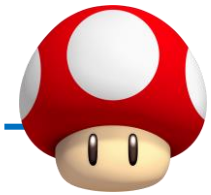
2 변수선택법

변수 선택 방법 | ① Best Subset Selection

Best Subset Selection

가능한 모든 변수들의 조합을 다 고려하는 방법

→ 변수의 개수가 p개라면, 2^p 개의 모형을 모두 적합하고 비교



▶ 가능한 모든 경우의 수를 고려하기 때문에 더 **신뢰할 수 있는 결과**를 산출

▶ p > 40인 경우 **계산 불가능**

▶ 적당한 p에서도 관측치가 많을 경우 **계산 비용**이 많이 소모

변수 선택 방법 | ① Best Subset Selection

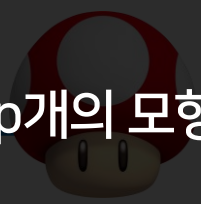
Best Subset Selection's Algorithm

Best Subset Selection

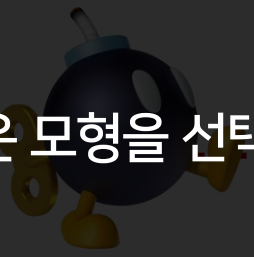
가능한 모든 변수들의 조합을 다 고려하는 방법

→ 변수의 개수가 p개라면, 2^p 개의 모형을 모두 적합하고 비교

1. $M_1, \cdots M_p$ 개의 모형을 적합한다. 이때 $M_k (k = 1,2, \ldots, p)$ 는 변수의 개수를 k개로 적합했을 때의 회귀식 중에 training error(주로 MSE)가 제일 작은 식이다.



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



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

더 신뢰할 수 있는 결과를 산출

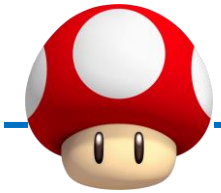
→ 큰 p에서도 관측치가 많을 경우

계산 비용이 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 매우 빠름**
- ▶ **변수의 개수가** 관측치의 개수보다 **많은 경우**에도 사용 가능

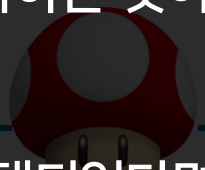
- ▶ 가능한 모든 변수 조합을 고려하지는 않기 때문에 **최적의 모형으로 확신할 수 없음** (Local Optimum)

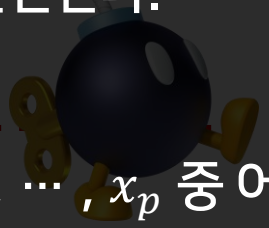
변수 선택 방법 | ② 전진선택법

Forward Selection's Algorithm

전진선택법 *Forward Selection*

1. 상수항만을 포함하고 있는 Null Model($y = \beta_0$)에서 시작해 $x_1, \cdots, x_p$ 중에 어떤 변수를 추가하는 것이 AIC 또는 BIC를 낮추는지 판단한다.





2. 만일 x_1이 선택되었다면, $y = \beta_0 + \beta_1 x_1$ 식에서 $x_2, \cdots, x_p$ 중 어떤 변수를 추가하는 것이 AIC 또는 BIC를 낮추는지 판단한다.

계산이 매우 빠름

가능한 모든 변수 조합을
고려하지는 않기 때문에

(Local Optimum)

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

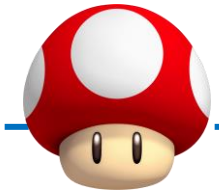
변수 선택 방법 | ③ 후진제거법

후진제거법 *Backward Elimination*

Full Model$\left(y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\right)$ 에서 시작해,

변수를 하나씩 제거하는 방법

Forward Selection의 반대!



▶ Best Subset Selection에 비해
계산이 매우 빠름

▶ Best Subset Selection 처럼
p > 40인 경우 **계산 불가능**

▶ **최적의 모형으로 확신할 수 없음**
(Local Optimum)

2 변수선택법

변수 선택 방법 | ③ 후진제거법

Backward Elimination's Algorithm

후진제거법 *Backward Elimination*

Full Model$\left(y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\right)$ 에서 시작해,

1. Full Model에서 시작해 $x_1, \cdots, x_p$ 중에 AIC 또는 BIC를 가장 크게 낮추는 변수를 선택하여 제거한다.

2. 이러한 과정을 반복하며 AIC와 BIC가 낮아지면 계속 제거하고, 더 이상 AIC와 BIC가 낮아지지 않는다면 프로세스를 중단한다.

▶ Best Subset Selection에 비해 계산이 매우 빠름

▶ Best Subset Selection 처럼 $p > 40$인 경우 계산 불가능

▶ 최적의 모형으로 확신할 수 없음 (Local Optimum)

변수 선택 방법 | ④ 단계적 선택법

단계적 선택법

Forward Selection과 Backward Elimination 과정을 **섞은 방법**

Null model 혹은 Full Model에서 시작하지만,
변수를 선택 및 제거하는 경우를 모두 고려했을 때
AIC와 BIC가 **감소하는 방향**으로 움직임

변수 선택 방법 | ④ 단계적 선택법

Stepwise Selection's Algorithm 기본형

단계적 선택법

Forward Selection과 Backward Elimination 과정을 섞은 방법

1. Null model 혹은 Full model에서 시작한다.

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

Null model 혹은 Full Model에서 시작하지만,
변수를 한 개씩 제거하는 경우를 모두 고려했을 때
AIC와 BIC가 감소하는 방향으로 움직임

3. AIC, BIC가 가장 작은 모형을 선택한다.

변수 선택 방법 | ④ 단계적 선택법

Stepwise Selection's Algorithm 예시

단계적 선택법

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.

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

2 변수선택법

변수 선택 방법 | ④ 단계적 선택법

Stepwise Selection's Algorithm 예시

단계적 선택법

Forward Selection과 Backward Elimination 과정을 섞은 방법

전진선택법을 사용할 때 한 변수가 선택되면,

이미 선택된 변수 중 중요하지 않은 변수가 있을 수 있음

Null model 혹은 Full model에서 시작하지만,

변수를 선택 및 제거하는 경우를 모두 고려했을 때

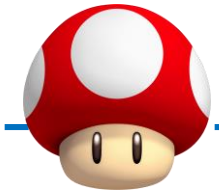
전진선택법의 각 단계에서 이미 선택된 변수들의 중요도를

다시 검사하여 중요하지 않은 변수를 제거하는 방법

변수 선택 방법 | ④ 단계적 선택법

단계적 선택법

Forward Selection과 Backward Elimination 과정을 **섞은 방법**



▶ Best Subset Selection에 비해 계산이 매우 빠름

▶ 변수 제거와 추가 모두 가능하다는 점에서 상대적으로 유연하나, **모든 변수 조합**을 고려하는 것은 아니므로 Best Model이라 할 수는 **없음**

2 변수선택법

변수 선택 방법 | ④ 단계적 선택법

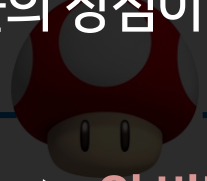
단계적 선택법

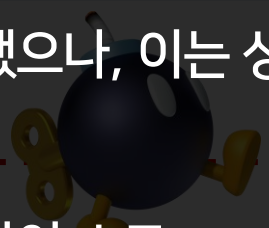
정리하자면,

Forward Selection과 Backward Elimination 과정을 **섞은 방법**

▶ Best Subset Selection을 제외한

나머지 방법들의 장점이 **계산이 매우 빠른 것**이라고 했으나, 이는 상대적인 것





▶ **위 방법 모두** 계산 비용이 굉장히 많이 소모

Best Subset Selection에 비해
계산이 매우 빠름

변수 제거와 추가 모두 가능하다는
점에서 상대적으로 유연하나, **모든**
변수 조합을 고려하는 것은 아니므로
Best Model이라 할 수는 **없음**

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

2 변수선택법

변수 선택 방법 | ④ 단계적 선택법

단계적 선택법

정리하자면,

Forward Selection과 Backward Elimination 과정을 섞은 방법

기계적으로 변수를 추가 혹은 제거하는 행위는 매우 위험

정규화 방법!

Best Subset Selection에 비해 계산이 매우 빠름

변수 제거와 추가 모두 가능하다는 점에서 상대적으로 유연하나, **모든 변수 조합**을 고려하는 것은 아니므로 Best Model이라 할 수는 **없음**

3

정규화

3 정규화

정규화

정규화

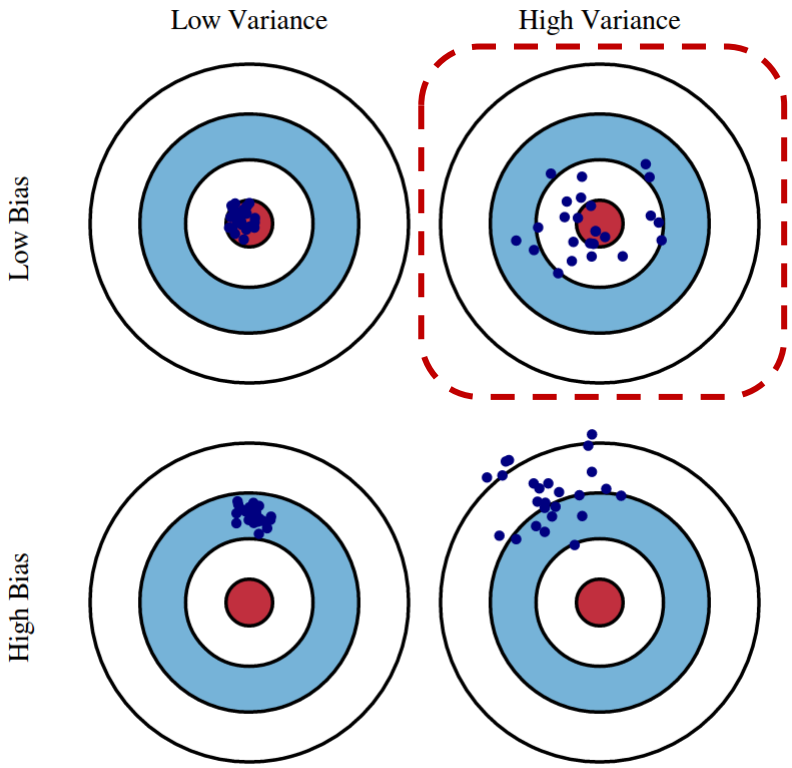
회귀계수가 가질 수 있는 값에 **제약조건**을 부여하여
계수들을 작게 만들거나 0으로 만드는 방법

다중공선성은 OLS 추정량의 **분산을 크게 증가**시킴

정규화는 OLS 추정량의 **불편성**을 포기하고,
분산을 줄임

3 정규화

Bias-Variance Trade off



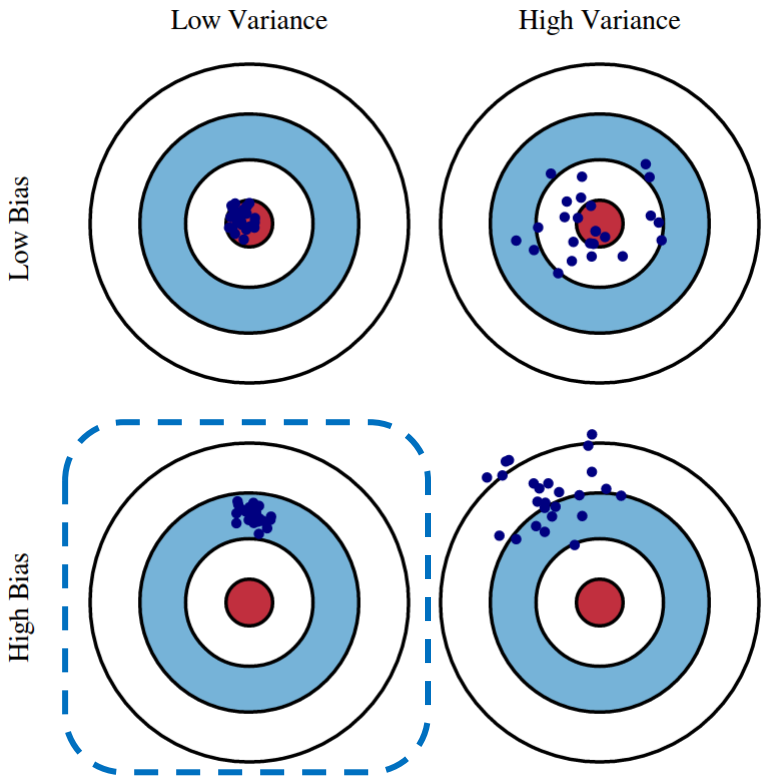
정규화 이전

다중공선성이 존재하는 경우

회귀계수 OLS 추정량

3 정규화

Bias-Variance Trade off



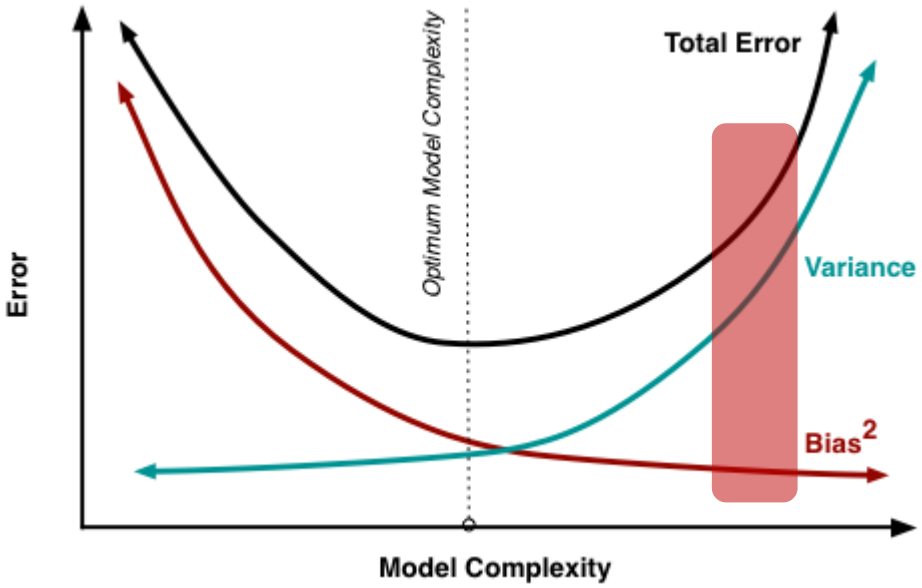
정규화 이후

정규화로 불편성을 포기하고

분산을 줄이는 효과

3 정규화

Bias-Variance Trade off



정규화 이전,

다중공선성이 존재하는 경우

편향이 낮고 분산이 높음

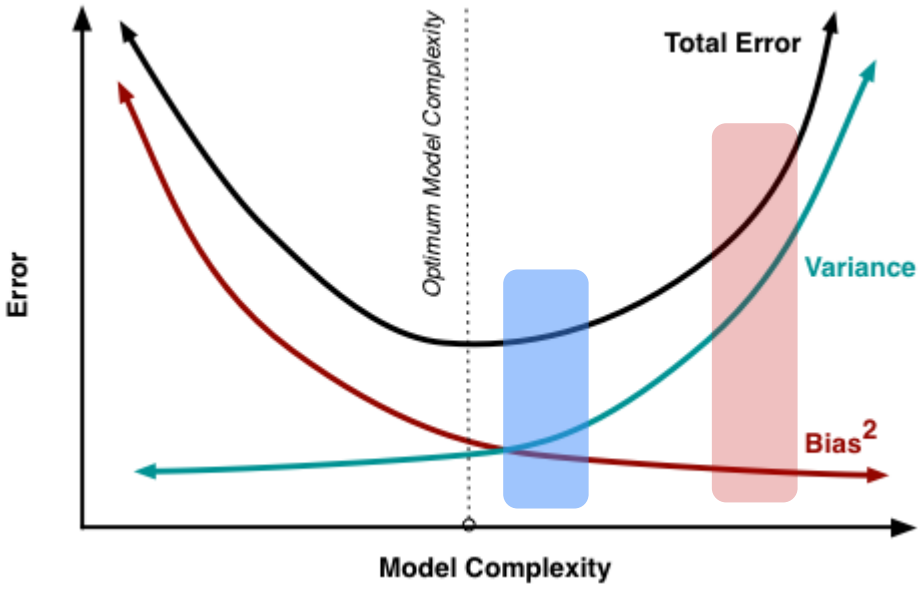
정규화 이후,

편향은 중가하지만

분산이 크게 줄어듦

3 정규화

Bias-Variance Trade off



정규화 이전,

다중공선성이 존재하는 경우

편향이 낮고 분산이 높음

⋮

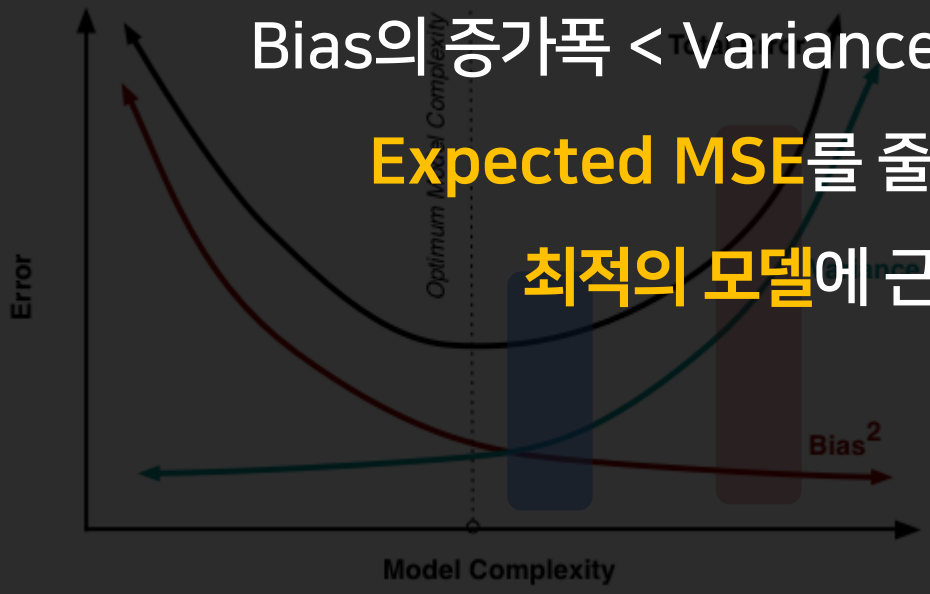
정규화 이후,

편향은 증가하지만

분산이 크게 줄어들음

3 정규화

Bias-Variance Trade off



Bias의 증가폭 < Variance의 감소폭을 만족해

Expected MSE를 줄일 수만 있다면

최적의 모델에 근접 가능!

정규화 이전,

다중공선성이 존재하는 경우

편향이 낮고 분산이 높음

정규화 이후,

편향은 증가하지만

분산이 크게 줄어들음

정규화 | 목적함수

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 (정규화의 증거)

3 정규화

정규화 | 목적함수

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 (정규화의 증거)

λ는 우리가 조절할 수 있는 **하이퍼파라미터**로

LSE와 Generalization Accuracy 사이의 **Trade-off**를 조절하는 역할을 수행

정규화 | 목적함수

목적함수

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

예시

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

위 식과 같이 β_3^2와 β_4^2의 계수가 크다면,

Expected MSE를 최소로 하기 위해 $\beta_3 \approx 0$, $\beta_4 \approx 0$이 되어야 함

3 정규화

목적함수

목적함수

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

▶ λ**가 매우 크다면**, $\beta_1 \approx 0, \beta_2 \approx 0, \beta_3 \approx 0, \beta_4 \approx 0 \rightarrow y = \beta_0$(직선)

▶ λ**가 매우 작다면**, β 에 대한 제약이 거의 없는 것

정규화 | ① Ridge

Ridge *(L2 Regularization)*

SSE를 최소화하면서 회귀계수 β**에 제약조건**을 거는 방법

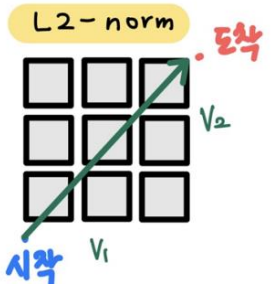
제약 조건식이 **L2-norm** 형태

⋮

L2 Regularization으로 불리는 이유?

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

제약조건식이 L2-norm 형태이기에 L2 Regularization으로 불림



정규화 | ① Ridge

목적함수

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

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

⋮

목적함수를 최소화함으로써 Ridge Estimator 추정

β**에 대한 이차식** 형태이므로 **미분**을 통해 추정량 계산 가능

정규화 | ① Ridge

목적함수

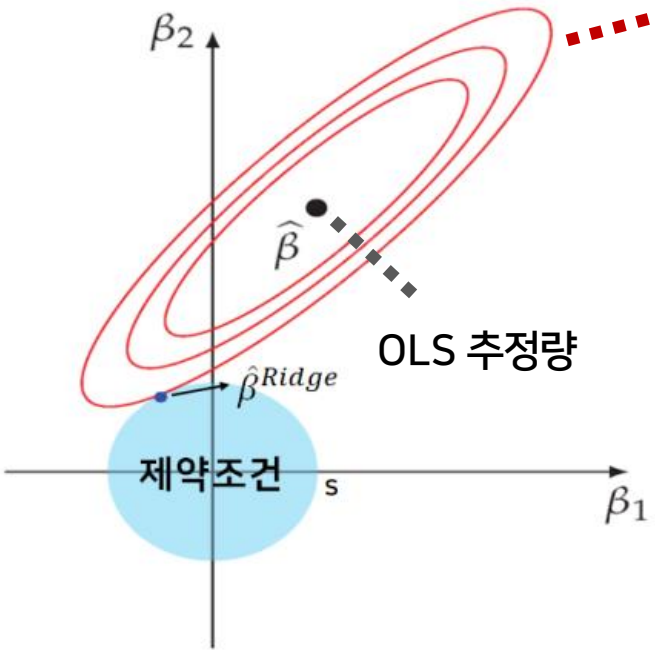
$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j \right)^2 \; subject\ to \sum_{j=1}^{p} \beta_j^2 \leq s$$

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

단, **설명 변수**들은 **표준화된 상태**여야 함

정규화 | 목적함수에 대한 이해 ①

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



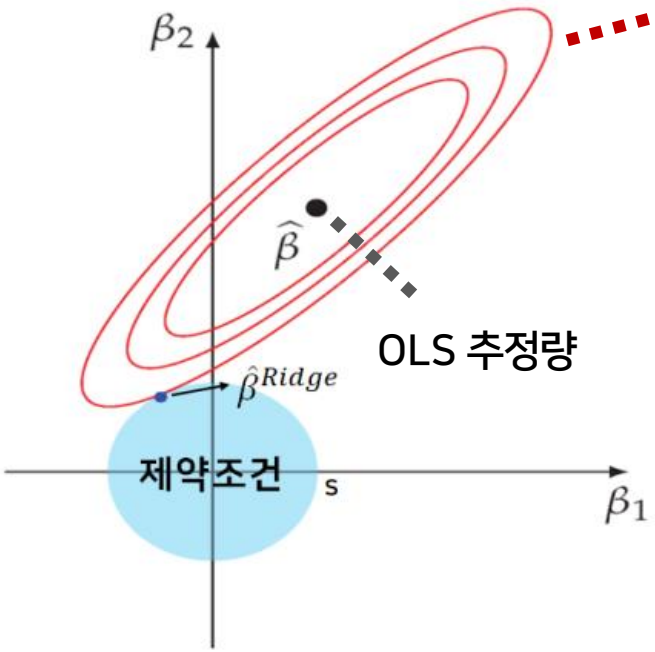
회귀계수의 최소화

① 회귀계수 $\hat{\beta}$는 반드시 원 내부에 존재

② SSE를 최소화해야 함

정규화 | 목적함수에 대한 이해 ①

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j \right)^2 \; subject\ to \sum_{j=1}^{p} \beta_j^2 \le s$$



추정량의 SSE

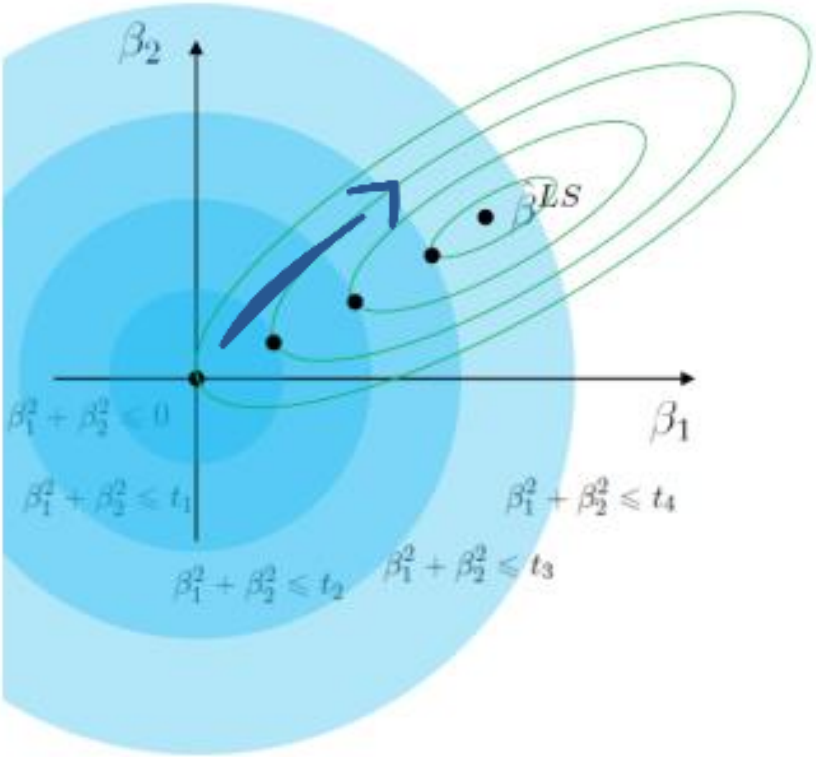
OLS 추정량

회귀계수의 최소화

타원과 원의 접점이 곧

Ridge Estimator

정규화 | 목적함수에 대한 이해 ①



제약조건이 **완화**될 경우

s가 증가하면서 원의 넓이 증가

원이 타원을 밀어내며 추정량이 0에서 멀어짐

회귀계수를 **작게 만들 수 없음**

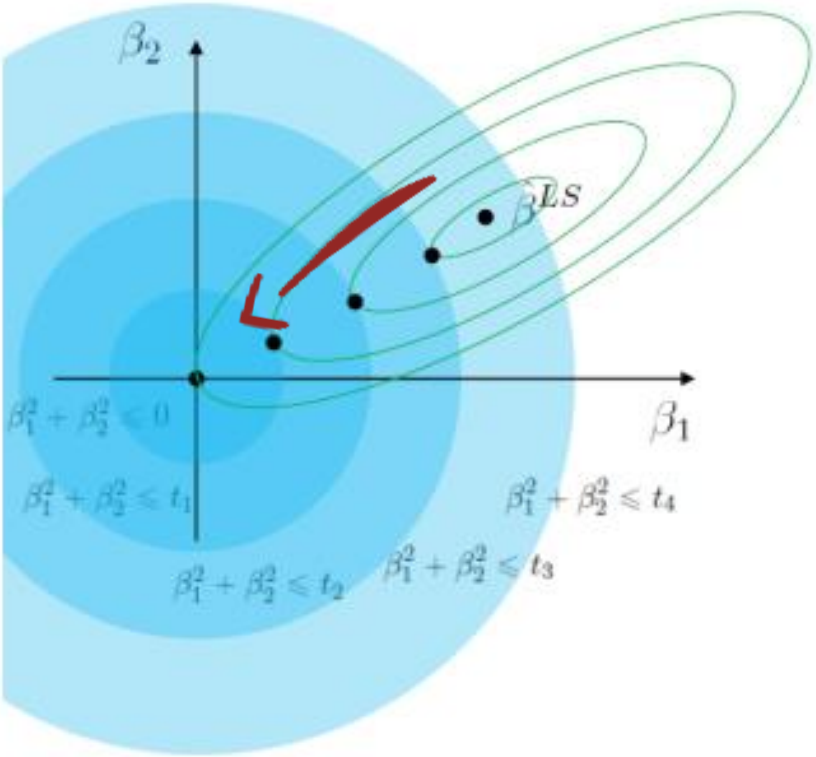
제약조건이 **강화**될 경우

s가 감소하면서 원의 넓이 감소

추정량이 0으로 수렴함(**0은 될 수 없음**)

회귀계수를 **작게 만들 수 있음**

정규화 | 목적함수에 대한 이해 ①



제약조건이 완화될 경우

s가 증가하면서 원의 넓이 증가

원이 타원을 밀어내며 추정량이 0에서 멀어짐

회귀계수를 작게 만들 수 없음

제약조건이 강화될 경우

s가 감소하면서 원의 넓이 감소

추정량이 0으로 수렴함(0은 될 수 없음)

회귀계수를 작게 만들 수 있음

정규화 | 목적함수에 대한 이해 ②

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

라그랑주 승수법을 이용해 나타낸 함수식

⋮

오차제곱합(SSE) 최소화 & **Regularization term**을 통해

개별 회귀계수 크기 조정

정규화 | 목적함수에 대한 이해 ②

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

λ는 음수가 아닌 Tuning Parameter

최적의 모델을 찾는 과정에서 **직접 CV를 통해 조정**해주는 모수로 ,
제약조건의 크기를 결정 (s와는 반대 관계)

정규화 | 목적함수에 대한 이해 ②

λ의 값에 따른 회귀계수의 변화

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

λ가 커지는 경우

λ의 영향력이 증가하므로,

전체 식을 최소화하기 위해

$\sum_{j=1}^{p} \beta_j^2$은 작아져야 함

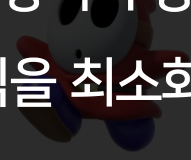
▶ 개별 회귀 계수들은 감소

λ가 작아지는 경우

λ의 영향력이 감소하므로,

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

▶ 개별 회귀 계수들은 증가



λ는 음수가 아닌 Tuning Parameter

최적의 모델을 찾는 과정에서 직접 CV를 통해 조정해주는 모수로,

제약조건의 크기를 결정 (s와는 반대 관계)

3 정규화

정규화 | 목적함수에 대한 이해 ②

λ의 값에 따른 회귀계수의 변화

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

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

λ는 함수가 아닌 Tuning Parameter

최적의 λ를 찾는 과정에서 직접 CV를 통해 조정해주는 모수로 ,

제약조건의 크기를 결정 (s와는 반대 관계)

Ridge | 특징

Scaling

회귀계수는 **변수 단위**에 큰 영향을 받음

단위의 영향을 제거, 순수 영향력만 사용

주로 standard scaling 사용

계산 비용 절약

Regularization term이

이차식 형태이므로 **미분 가능**

λ를 바꾸며 미분과 함께 행렬 연산

예측 성능

상관관계가 높은 변수들이 모델에

존재할 경우, 좋은 예측 성능을 보임

변수 선택

영향력을 줄일 뿐 변수는 잔존

다중공선성을 일으키는 **변수 제거 불가**

해석력 증가는 **기대하기 어려움**

Ridge | 특징

Scaling

회귀계수는 **변수 단위**에 큰 영향을 받음

단위의 영향을 제거, 순수 영향력만 사용

주로 standard scaling 사용

계산 비용 절약

Regularization term이

이차식 형태이므로 **미분 가능**

λ를 바꾸며 미분과 함께 행렬 연산

예측 성능

상관관계가 높은 변수들이 모델에

존재할 경우, 좋은 예측 성능을 보임

변수 선택

영향력을 줄일 뿐 변수는 잔존

다중공선성을 일으키는 **변수 제거 불가**

해석력 증가는 기대하기 어려움

3 정규화

Ridge | 특징

Scaling

회귀계수는 **변수 단위**에 큰 영향을 받음

단위의 영향을 제거, 순수 영향력만 사용

주로 standard scaling 사용

계산 비용 절약

Regularization term이

이차식 형태이므로 **미분 가능**

λ를 바꾸며 미분과 함께 행렬 연산

예측 성능

상관관계가 높은 변수들이 모델에

존재할 경우, **좋은 예측 성능**을 보임

변수 선택

영향력을 줄일 뿐 변수는 잔존

다중공선성을 일으키는 **변수 제거 불가**

해석력 증가는 기대하기 어려움

3 정규화

Ridge | 특징

Scaling

회귀계수는 **변수 단위**에 큰 영향을 받음

단위의 영향을 제거, 순수 영향력만 사용

주로 standard scaling 사용

계산 비용 절약

Regularization term이

이차식 형태이므로 **미분 가능**

λ를 바꾸며 미분과 함께 행렬 연산

예측 성능

상관관계가 높은 변수들이 모델에

존재할 경우, **좋은 예측 성능**을 보임

변수 선택

영향력을 줄일 뿐 변수는 잔존

다중공선성을 일으키는 **변수 제거 불가**

해석력 증가는 **기대하기 어려움**

3 정규화

Ridge | 특징

행렬연산을 통한 Closed Form Solution

Scaling

회귀계수는 **변수 단위**에 영향을 받음

단위의 영향을 제거, 순수 영향력만 사용

주로 standardized 사용

계산 비용 절약

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

tion term이

이차식 형태이므로 **미분 가능**

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

을 한꺼번에 행렬 연산

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

예측 성능

상관관계가 높은 변수들이 모델에

존재할 경우, 좋은 예측 성능을 보임

변수 선택

영향력을 줄일 뿐 변수는 잔존

다중공선성을 일으키는 **변수 제거 불가**

해석력 증가는 기대하기 어려움

$\boldsymbol{I_p}$는 $p \times p$ 크기의 Identity Matrix이므로

대각요소 각각에 $\boldsymbol{\lambda}$ 만큼 더해주었다고 이해 가능

3 정규화

Ridge | 특징

행렬연산을 통한 Closed Form Solution

Scaling

회귀계수는 **변수 단위**에 큰 영향을 받음

단위의 영향을 제거, 순수 변수 간 영향력만 비교

주로 standard scaling 사용

계산 비용 절약

Regularization term이

어떠한 상태이므로 **미분 가능**

λ 로 바꾸면 미분과 함께 행렬 연산

LS는 BLUE에 의해 Unbiased

Ridge는 λ 만큼 더했기에 Biased

예측 성능

상관관계가 높은 변수들이 많이

존재할 경우, **좋은 예측 성능**을 보임

변수 선택

그러나 **Variance가 작기** 때문에

더 높은 예측 성능을 가짐

영향력을 줄일 뿐 변수는 잔존

다중공선성을 일으키는 **변수 제거 불가**

해석력 증가는 기대하기 어려움

3 정규화

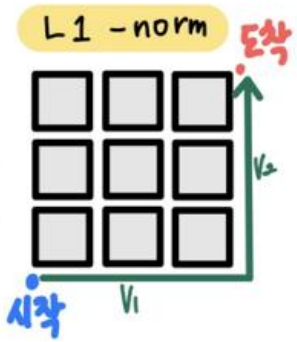
Lasso

Lasso *(L1 Regularization)*

SSE를 최소화하면서 회귀계수 β에 제약을 거는 방법

Ridge와 같은 아이디어를 바탕으로 함

L1-norm



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

원점에서 벡터까지의 각 좌표의 합

3 정규화

Lasso | 목적함수

목적함수

$$\hat{\beta}^{Lasso} = \underset{\beta}{arg\min} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_i)^2 \ subjuect\ to \sum_{j=1}^{p} |\beta_j| \leq s$$

$$\vdots$$

$$\hat{\beta}^{Lasso} = \underset{\beta}{arg\min} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

위 식을 최소화함으로써 회귀계수의 Lasso Estimator를 얻을 수 있음

- ✓ 단, 미분이 불가능하므로 수치적인 방법을 이용하여 최적화 문제를 해결해야 함
- ✓ 설명 변수들은 표준화된 상태여야 함

3 정규화

Lasso | 목적함수

목적함수에서 s와 λ의 기능

목적함수

s 와 λ 는 정규화를 위한 제약조건이라는 점에서는 같지만

그 영향은 반대 방향으로 작용

$$\hat{\beta}^{Lasso} = \underset{\beta}{arg\min} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_i)^2 \quad subject\ to \sum_{j=1}^{p} |\beta_j| \leq s$$

s 가 작음 = λ가 큼 = 제약을 많이 가함

$$\hat{\beta}^{Lasso} = \underset{\beta}{arg\min} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_i x_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

s 가 큼 = λ가 작음 = 제약을 적게 가함

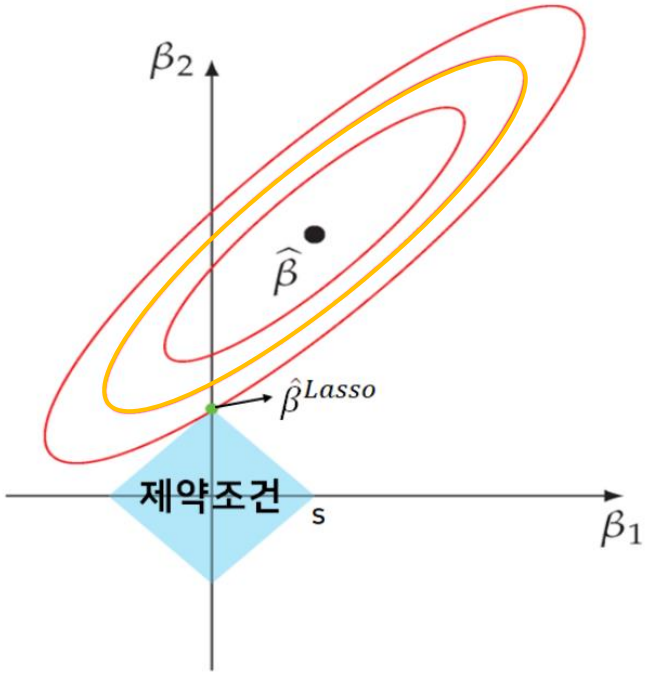
위 식을 최소화하는 곳에서 회귀계수의 [illegible]를 얻을 수 있음

- ✓ 단, 미분이 불가능하므로 수치적인 방법을 이용하여 최적화 문제를 해결해야 함
- ✓ 설명 변수들은 표준화된 상태여야 함

3 정규화

Lasso | 목적함수의 이해

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



빨간 타원

목적함수 SSE가 만드는 도형

파란 마름모

제약조건 $\sum_{j=1}^{p} |\beta_j| \le s$가 만드는 도형

⋮

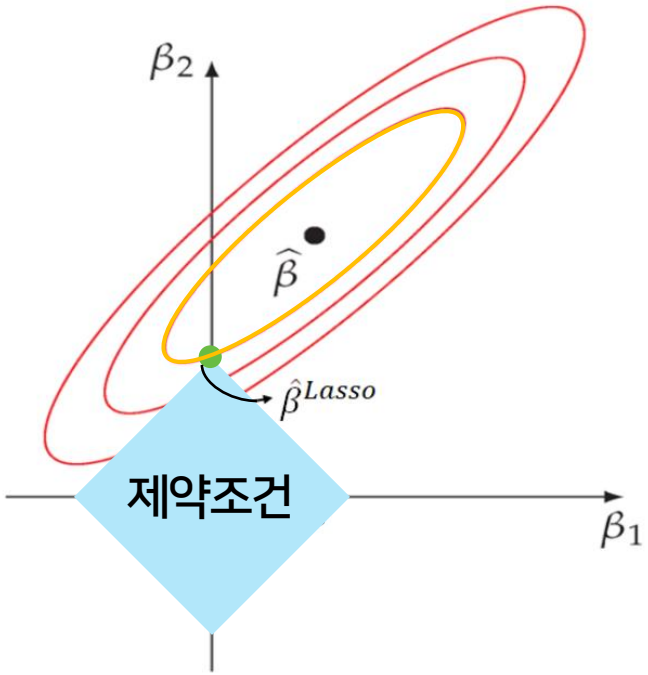
타원과 마름모의 접점이

회귀계수의 Lasso Estimator

3 정규화

Lasso | 목적함수의 이해

① s가 커질 때



마름모의 넓이가 커짐

⋮

제약 조건이 완화됨

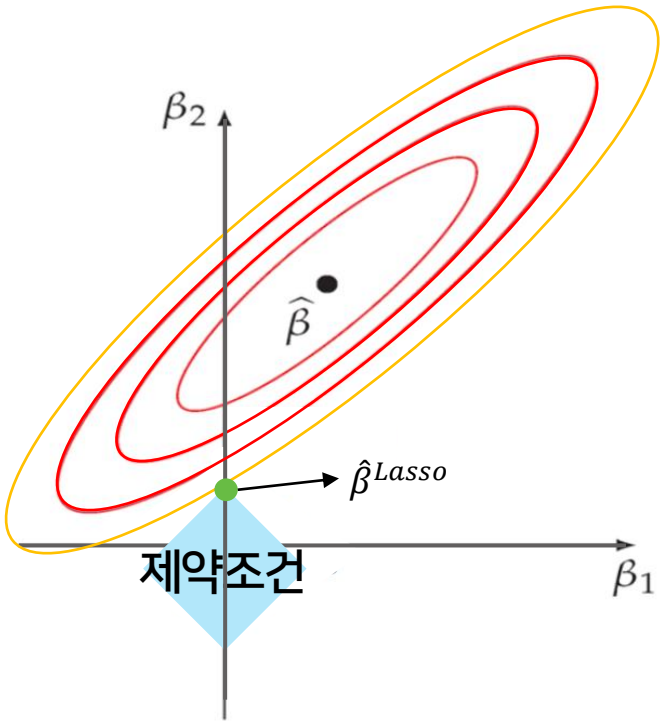
⋮

회귀계수 추정량이 0에서 멀어짐

3 정규화

Lasso | 목적함수의 이해

① s가 작아질 때



마름모의 넓이가 작아짐

⋮

제약 조건이 강화됨

⋮

회귀계수 추정량이 0에 가까워짐

3 정규화

Lasso | 목적함수의 이해

Ridge vs. Lasso

① s가 작아질 때

앞서 살펴본 Ridge와 Lasso는 매우 유사한 원리

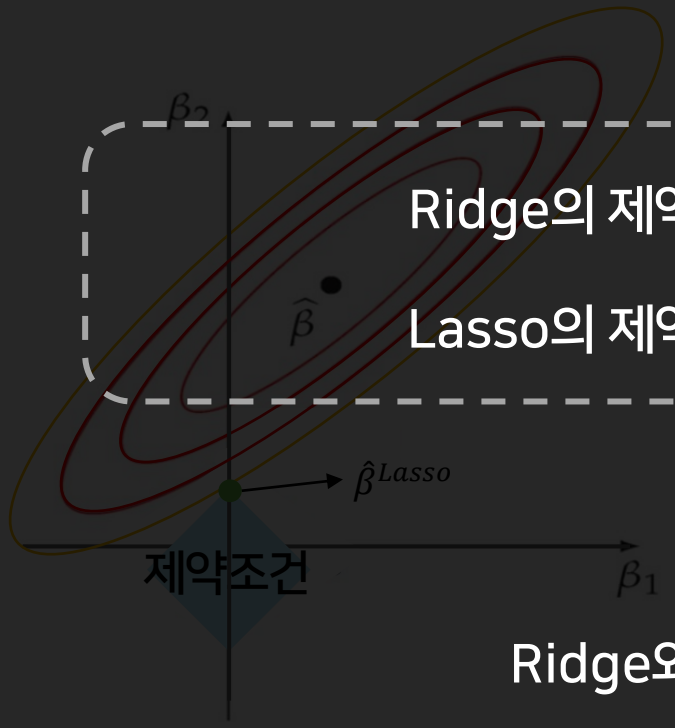
BUT!

마름모의 넓이가 작아짐

Ridge의 제약조건 : $\sum_{j=1}^{p} \beta_j{}^2 \leq s$

Lasso의 제약조건 : $\sum_{j=1}^{p} |\beta_j| \leq s$

제약 조건이 강화됨



회귀계수 추정량이 0에 가까워짐

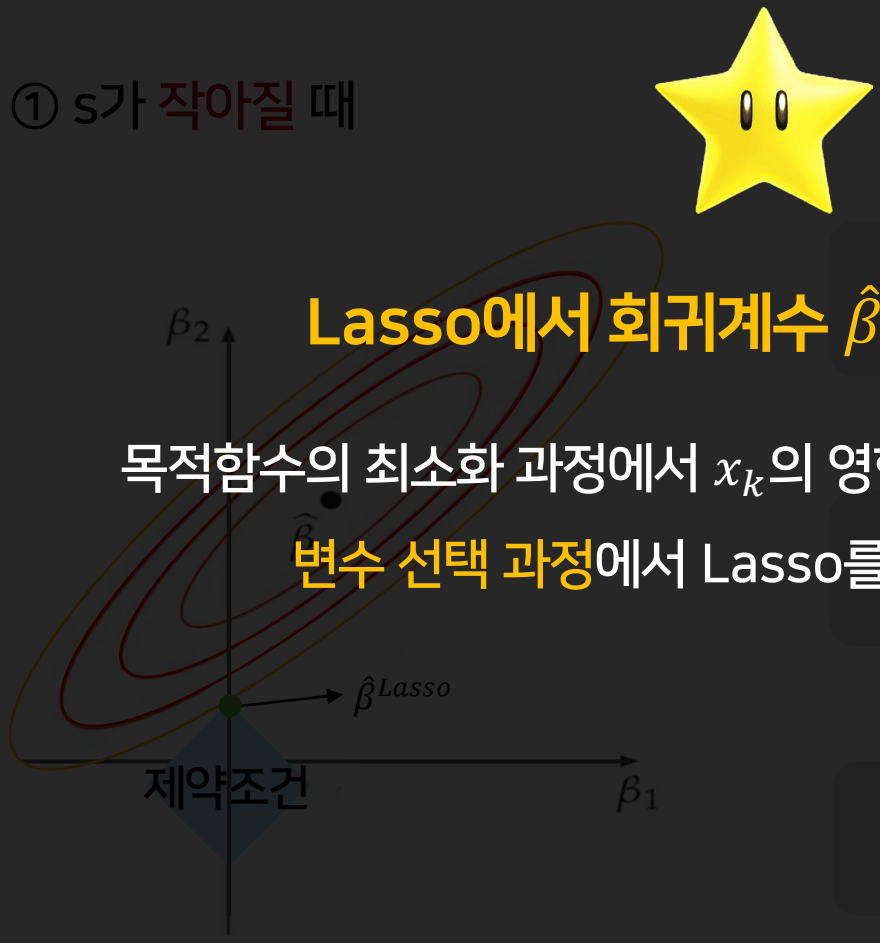
Ridge와 달리 Lasso는

일부 회귀계수 $\hat{\beta}$가 0이 되는 추정량이 도출될 수 있음

3 정규화

Lasso | 목적함수의 이해

① s가 작아질 때



Lasso에서 회귀계수 $\hat{\beta}_k$가 0이라면?

목적함수의 최소화 과정에서 x_k의 영향력이 사라지는 것이므로

변수 선택 과정에서 Lasso를 활용할 수 있음!

3 정규화

Lasso 목적함수의 이해

목적함수

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

⋮

라그랑주 승수법으로 변환된 목적함수

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

오차제곱합(SSE) term

개별 회귀계수가 너무 많아지는 것을 조정하는 Regularization term

3 정규화

Lasso 목적함수의 이해

λ의 값에 따른 회귀계수의 변화

목적함수

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

λ가 커지는 경우

λ의 영향력이 증가하므로,

전체 식을 최소화하기 위해

$\sum_{j=1}^{p} |\beta_j|$ 은 작아져야 함

▶ 개별 회귀 계수들은 감소

λ가 작아지는 경우

λ의 영향력이 감소하므로,

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

▶ 개별 회귀 계수들은 증가

라그랑주 승수법으로 변환된 목적함수

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

오차제곱합(SSE) term

개별 회귀계수가 너무 많아지는 것을 조절하는 Regularization term

Ridge와 동일!

3 정규화

Lasso 목적함수의 이해

λ의 값에 따른 회귀계수의 변화

목적함수

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

λ가 커지는 경우

$\lambda \rightarrow \infty$

개별 회귀계수의 영향력은

무시될 만큼 작아짐

▶ 회귀 계수≈ 0

λ가 작아지는 경우

$\lambda = 0$

Regularization term이 없어짐

▶ OLS 추정량과 동일

오차제곱합(SSE) term

개별 회귀계수가 너무 많아지는 것을 조절하는 Regularization term

Ridge와 동일!

3 정규화

Lasso 목적함수의 이해

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

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 해석 가능성 증가

예측 성능

변수들 간 상관관계가 큰 경우,

유의미한 변수들을 0으로 만들 수 있음

Ridge보다 상대적으로 예측 성능 저하

Closed Form Solution

미분 불가능한 점이 존재해

Closed Form Solution을 못 구함

▶ 수치 최적화 방법 사용

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 해석 가능성 증가

예측 성능

변수들 간 상관관계가 큰 경우,

유의미한 변수들을 0으로 만들 수 있음

Ridge보다 상대적으로 예측 성능 저하

Closed Form Solution

미분 불가능한 점이 존재해

Closed Form Solution을 못 구함

▶ 수치 최적화 방법 사용

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 해석 가능성 증가

▶ 변수 간 상관 관계가 높다면 변수 선택 성능이 떨어짐

▶ 0이 되는 계수의 존재로 인해 sparsity(희박성)을 가짐

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 해석 가능성 증가

예측 성능

변수들 간 상관관계가 큰 경우,

유의미한 변수들을 0으로 만들 수 있음

Ridge보다 상대적으로 예측 성능 저하

Closed Form Solution

미분 불가능한 점이 존재해

Closed Form Solution을 못 구함

▶ 수치 최적화 방법 사용

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 해석 가능성 증가

예측 성능

변수들 간 상관관계가 큰 경우,

유의미한 변수들을 0으로 만들 수 있음

Ridge보다 상대적으로 예측 성능 저하

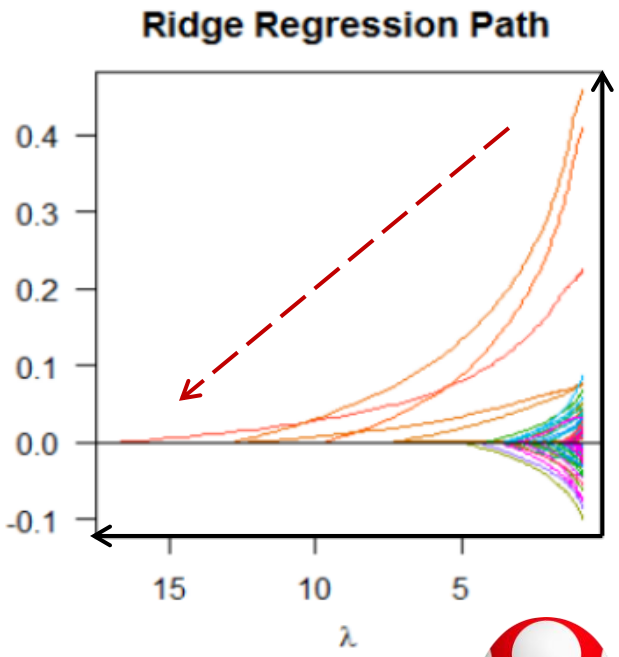
Closed Form Solution

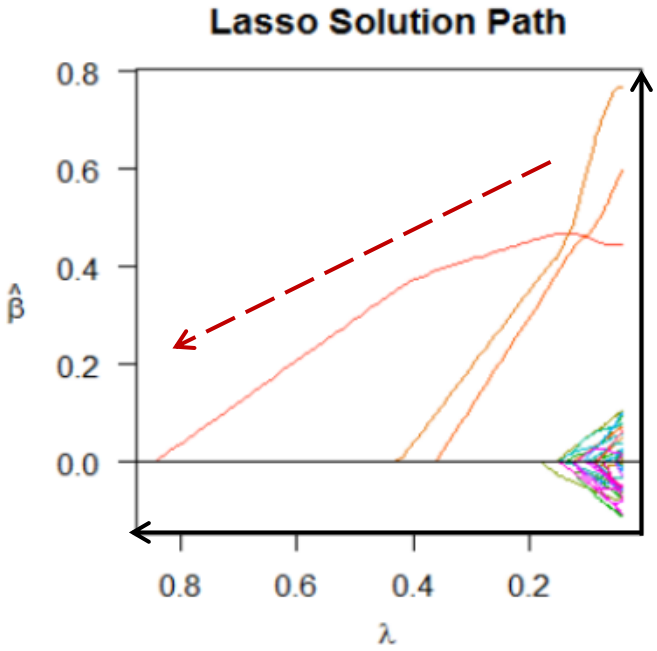
미분 불가능한 점이 존재해

Closed Form Solution을 못 구함

▶ 수치 최적화 방법 사용

Ridge vs. Lasso



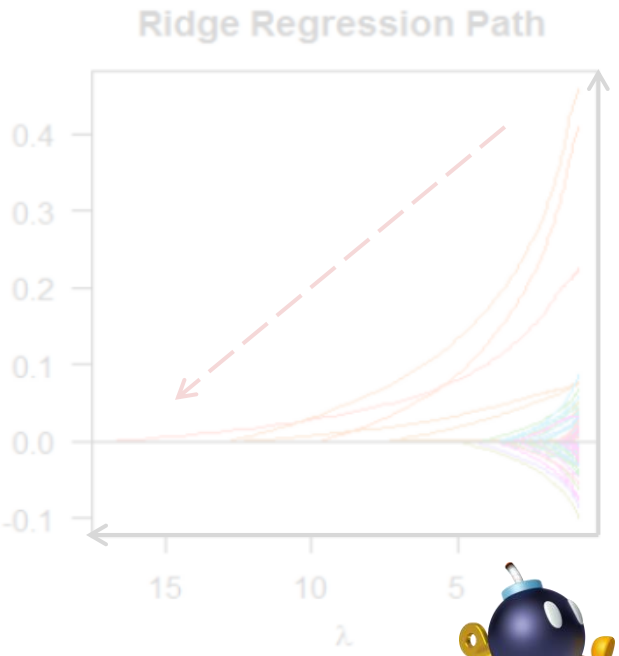


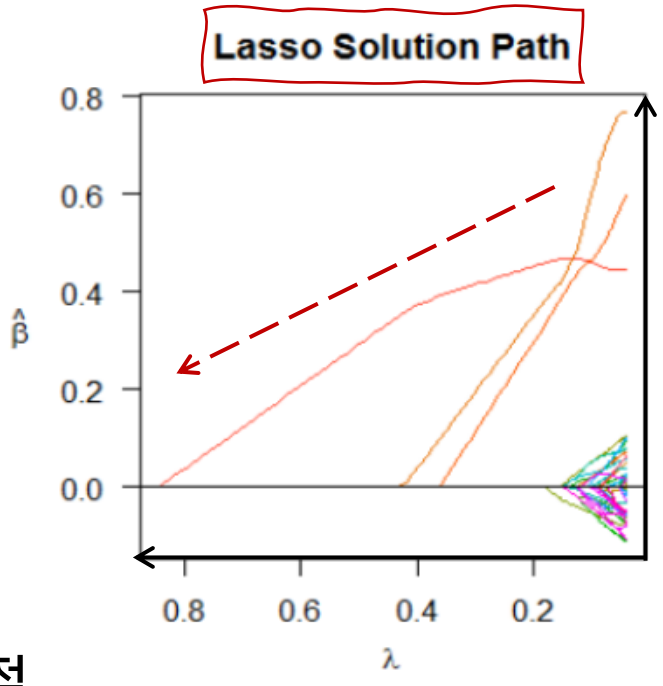
공통점

Ridge와 Lasso 모두 λ가 커짐에 따라(t가 작아짐에 따라) 모든 계수의 크기가 감소

3 정규화

Ridge vs. Lasso





차이점

- ▶ 중요하지 않은 변수가 더 빠르게 감소
- ▶ λ가 커짐에 따라(즉, t가 작아짐에 따라) 예측에 중요하지 않은 변수가 0이 됨

3 정규화

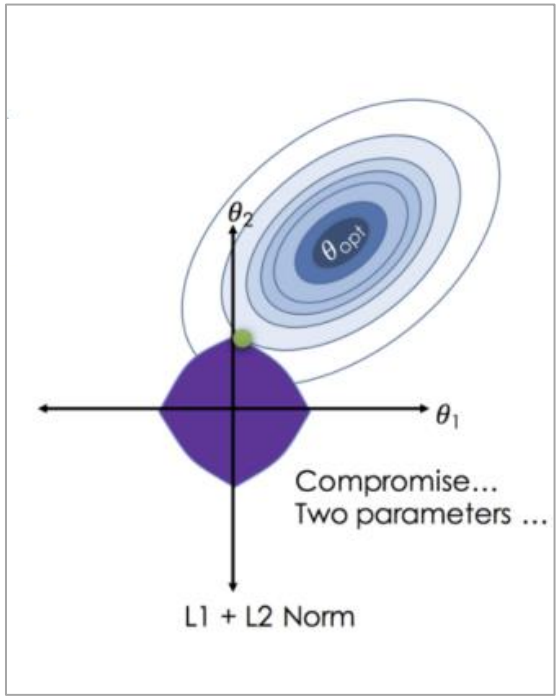
Ridge vs. Lasso | 정리

Ridge	Lasso
변수 선택 불가능	변수 선택 가능
Closed Form Solution O (미분이 가능함)	Closed Form Solution X (미분이 불가능함)
변수 간 상관관계가 높은 상황에서 좋은 예측 성능	변수 간 상관관계가 높은 상황에서 Ridge에 비해 예측 성능 ↓
제약 범위가 원	제약 범위가 마름모꼴
크기가 큰 변수를 우선적으로 줄임	

Elastic Net

Elastic Net

Ridge와 Lasso의 Regularization term을 혼합한 방법



Elastic Net의 제약조건은

Ridge, Lasso의 제약조건의 중간 형태

Elastic Net

Elastic Net

Ridge와 Lasso의 Regularization term을 혼합한 방법

변수 간 상관관계가 존재하는 경우

Lasso

상관관계가 존재하는 변수들 중
하나를 선택해 계수를 줄임

Elastic Net

상관관계가 존재하는 변수들을
모두 선택하거나 제거함

⋮

Grouping Effect에 의해 성능이 보완됨!

3 정규화

Elastic Net

Elastic Net

Grouping Effect

상관성이 있는 변수를 모두 선택 or 모두 제거해 성능 ↑

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

p_{ij} 는 x_i와 x_j의 상관계수

Lasso

상관관계가 존재하는 변수들 중 하나를 선택해 계수를 줄임

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

Elastic Net

상관관계가 존재하는 변수들을 모두 선택하거나 제거함

하나라도 중요하다면, 둘 다 똑같이 중요

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

Elastic Net | 목적함수

목적함수

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

$$subject\ to\ t_1 \sum_{j=1}^{p} |\beta_j| + t_1 \sum_{j=1}^{p} {\beta_j}^2 \leq s$$

$$\vdots$$

$$\hat{\beta}^{Elastic} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_i)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta^2$$

Lasso L1 term

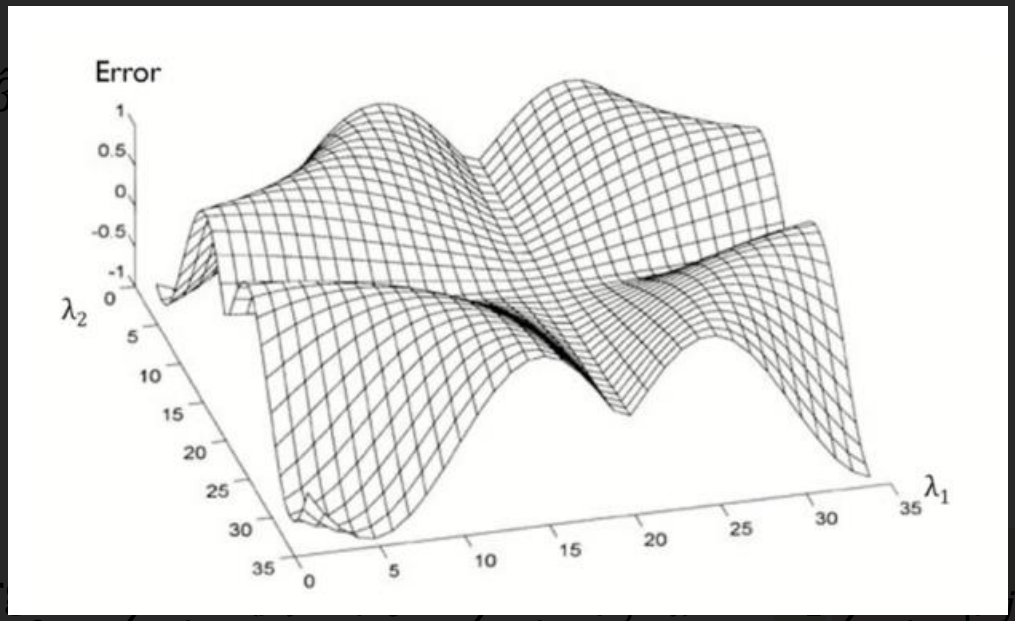
Ridge L2 term

3 정규화

Elastic Net | 목적함수

Elastic Net의 Parameter

목적함수



$$\hat{\beta}^{Elastic} = \arg\min_{\beta} \sum_{i=1}^{} \sum_{j=1}^{} \left| \sum_{j=1}^{} \right| + \lambda_2 \sum_{j=1}^{p} \beta^2$$

Grid Search 방법을 통해 범위 내에서

오차를 최소화하는 λ_1과 λ_2의 조합을 찾음

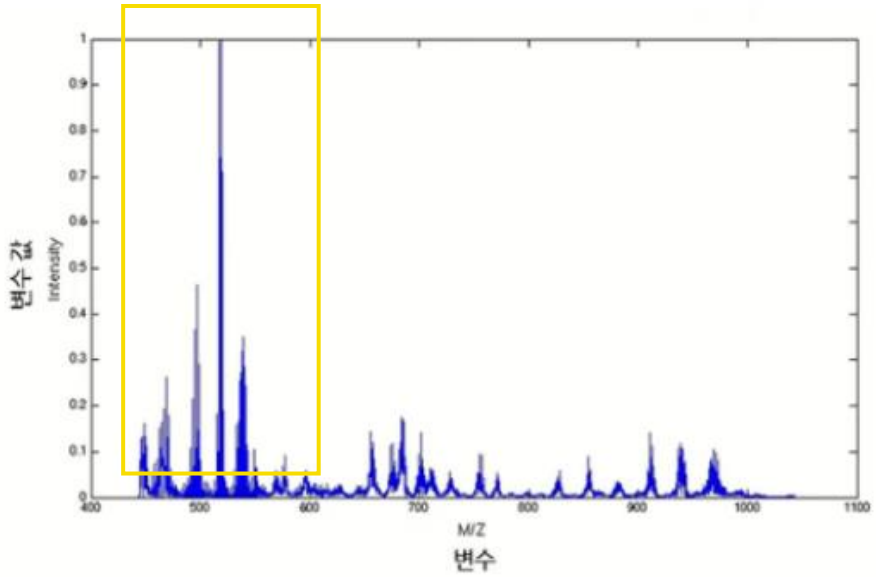
Ridge L2 term

3 정규화

Fused Lasso

Fused Lasso

변수들 간의 물리적인 거리에 대한 사전 지식을 이용한 정규화 모델



✓

중요한 변수들은 Peak를 기준으로

연속적으로 나타난다는 사실을 이용!

Fused Lasso

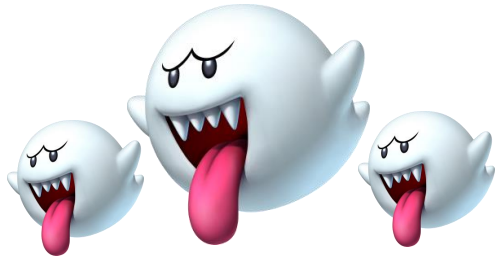
목적함수

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

Lasso L1 term

상관관계와 관계 없이 물리적으로 인접한 변수들의 회귀계수를 비슷한 값으로 추정하게 함

양 옆에 위치한 변수들의 회귀계수 값을 최소화하는 smoothness



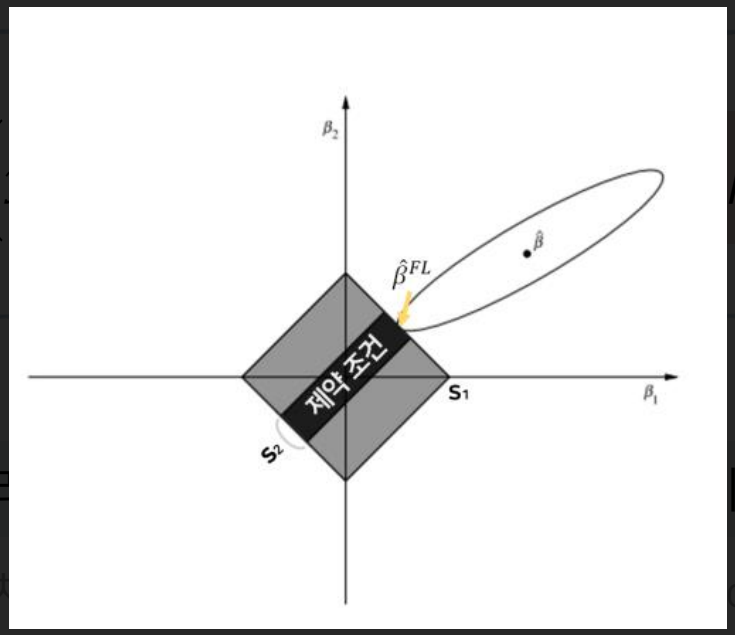
3 정규화

Fused Lasso

Fused Lasso 제약식의 시각화

목적함수

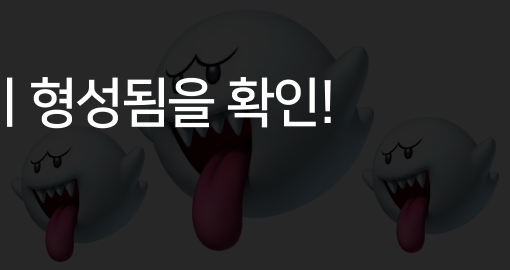
$$\hat{\beta}^{FL} = \underset{\beta}{\operatorname{argmin}}\left(\sum_{i=1}^{n}\left(\right.\right. \qquad \beta_j| + \lambda_2 \sum_{j=1}^{p} |\beta_j - \beta_j|)$$



상관관계와 관계 없이 물리 슷한 값으로 추정하게 함

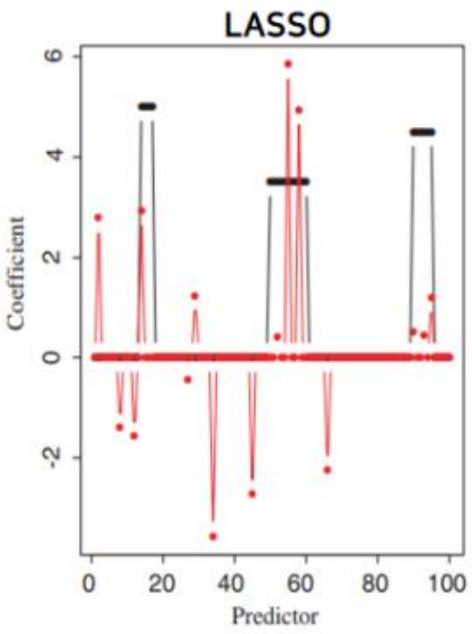
양 옆에 위치 oothness

기존 Lasso 제약 공간보다 더 엄격하게 제약 공간이 형성됨을 확인!

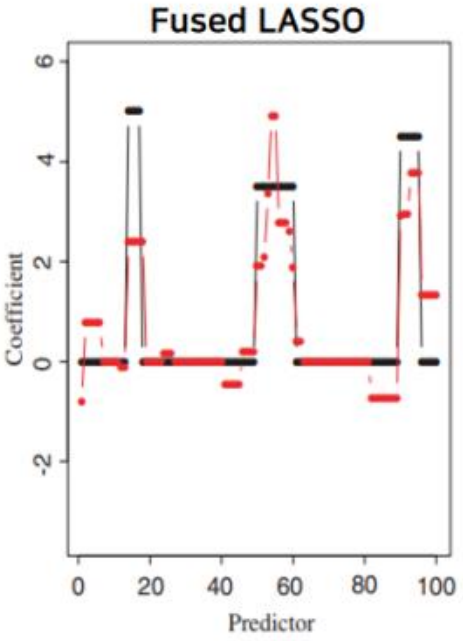


3 정규화

Fused Lasso



▲ 인접한 변수들의 실제 계수를 **잘 추정하지 못함**



▲ 인접한 변수들의 실제 계수를 **정확히** 추정

3 정규화

Adaptive Lasso

Adaptive Lasso

Lasso가 발전한 형태로,

각 회귀계수마다 다른 제약 조건을 주는 가중치 $\hat{w}_j$를 포함한 회귀모형

Lasso

모든 β_j에 동일한 제약을 주므로

β_j 자체가 크다면 회귀계수를 줄이는 데

최적화 과정이 Overfitting될 수 있음

Adaptive Lasso

큰 β_j에 작은 $\hat{w}_j$를 줌으로써

큰 회귀계수 최소화 과정에서의

Overfitting 문제 해결

3 정규화

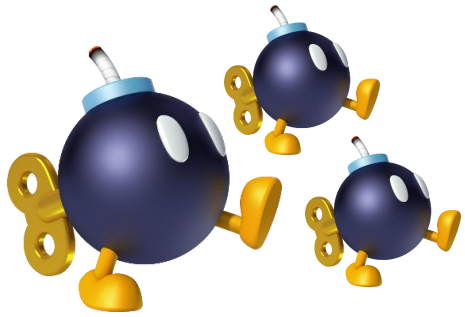
Adaptive Lasso

목적함수

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

$$\hat{w}_j = \frac{1}{(\hat{\beta}_j^{ini})^{\gamma}}$$

이때, $\hat{\beta}_j^{ini}$는 주로 Ridge Regression에 의한 X_j의 회귀계수



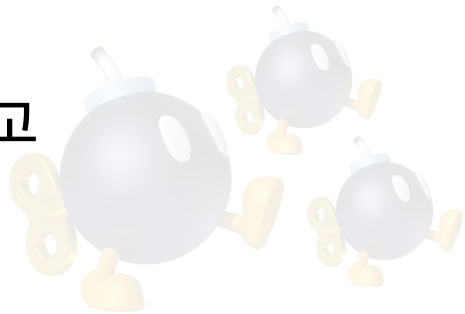
Adaptive Lasso

목적함수

$\hat{\beta}_j^{ini}$가 커짐	$\hat{\beta}_j^{ini}$가 작아짐
$\hat{w}_j$는 그에 반비례하여 작아짐	$\hat{w}_j$는 그에 반비례하여 커짐
β_j에 더 작은 제약이 가해짐	β_j에 더 큰 제약이 가해짐

특정 회귀계수로의 Overfitting을 방지하고

Oracle Property를 가지게 됨



3 정규화

Adaptive Lasso

Oracle Property

목적함수

$$\hat{\beta}^{AL} = \underset{\beta}{arg\min}\left(\sum \left(y_i - \beta_0 - \sum \beta_j x_j\right) + \lambda \sum_{j=1}^{p} \hat{w}_j |\beta_j|\right)$$

특정 모형에서 최적으로 선택된 변수가
항상 일관적으로 동일해야 하면서,
그 변수들로 fit된 모형의 예측도는 항상 최고여야 한다는 특성

$$\hat{w}_j = \frac{1}{(\hat{\beta}_j^{ini})^{\gamma}}$$

이때, $\hat{\beta}_j^{ini}$는 주로 Ridge Regression에 의한 X_j의 회귀계수

Lasso는 Oracle Property를 가지지 않지만,
Adaptive Lasso는 Oracle Property를 가짐

Thank you!