회귀분석팀

6팀
고경현
박세령
박이현
박지성
심예진
이선민

INDEX

0

REVIEW

잔차 플랏 출력

잔차 분포를 통해 **경험적 판단**에 근거한 **회귀진단** 가능

R에서 Plot() 함수를 통해 잔차의 분포를 표현

1) Residuals vs Fitted

2) Normal Q-Q plot

3) Scale – Location

4) Residuals vs Leverage

선형회귀분석의 가정

선형회귀분석의 기본 가정

선형성	정규성
독립성	등분산성

선형성 가정

선형성 가정

반응변수(Y)가 설명변수(X)의 선형결합으로 이루어짐

선형성 가정 ⇒ 단순선형회귀 다중선형회귀

선형성 가정이 위배되었다면?

변수 변환이나 비선형 모델 추정으로 대처 가능!

정규성 가정

정규성 가정

반응 변수 Y를 측정할 때 발생하는 오차는

정규분포를 따를 것이라는 가정

회귀식이 데이터를 잘 표현한다면!

잔차들은 단순한 측정 오차, 즉 Noise라 여겨짐

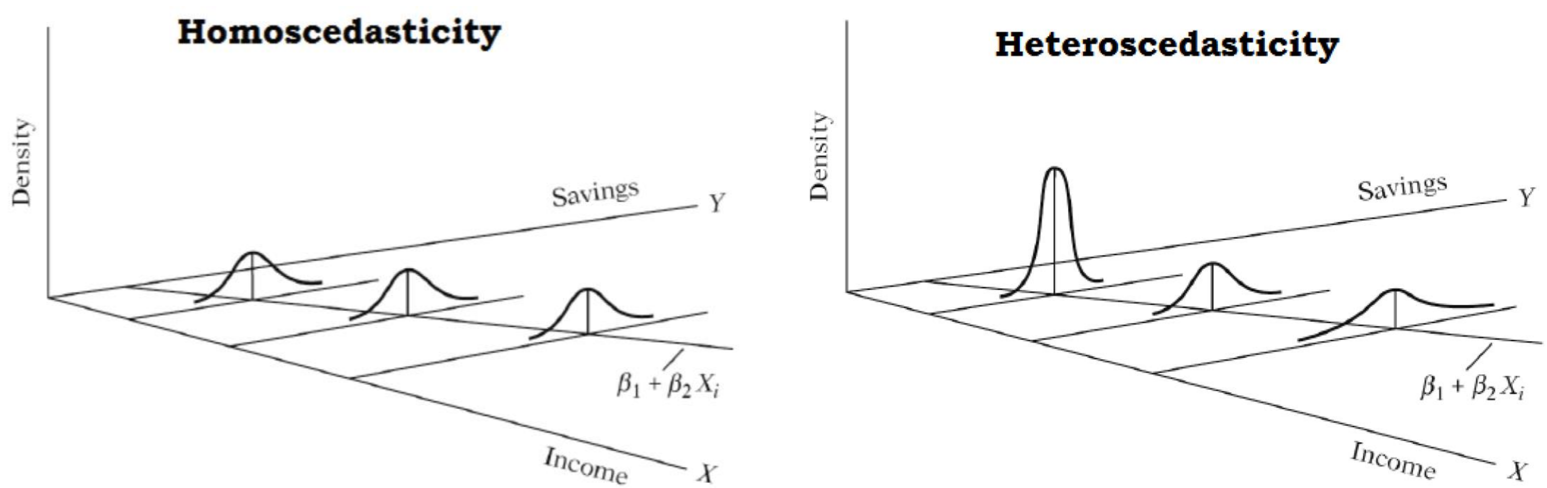
잔차들의 분포는 정규분포와 흡사한 형태

등분산성 가정

등분산성 가정

오차항의 분산이 어느 관측치에서나 동일하며

다른변수의 영향을 받지 않는 즉, **상수**라는 가정



독립성 가정

독립성 가정

$$Cov(e_i, e_j) = 0$$

오차들은 서로 독립이며, 개별 관측치에서 i번째 오차와
j번째 오차가 발생하는 것에 서로 영향을 미치지 않는다는 가정

독립성 가정이 위배된다면!

- 오차들간의 자기상관(autocorrelation)이 있다고 함
- 시공간 상의 데이터일 경우 오차들에 일종의 패턴이 존재할 수 있음

1

다중공선성

1 다중공선성

다중공선성이란?

설명변수 X_j들이 서로 선형적인 상관관계가 존재

설명변수가 서로 간의 선형결합으로 표현 가능

2주차 클린업 참고

변수에 대한 가정

선형성

설명변수들은 서로 독립

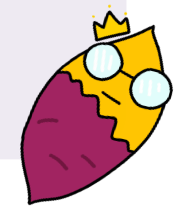
설명변수는 확률 변수 X

1 다중공선성

다중공선성이란?

설명변수 X_j들이 서로 선형적인 상관관계가 존재

설명변수가 서로 간의 선형결합으로 표현 가능



2주차 클린업 참고

변수에 대한 가정

선형성

설명변수들은 서로 독립

설명변수는 확률 변수 X

다중공선성이란?

설명변수 X_j들이 서로 **선형적인 상관관계**가 존재

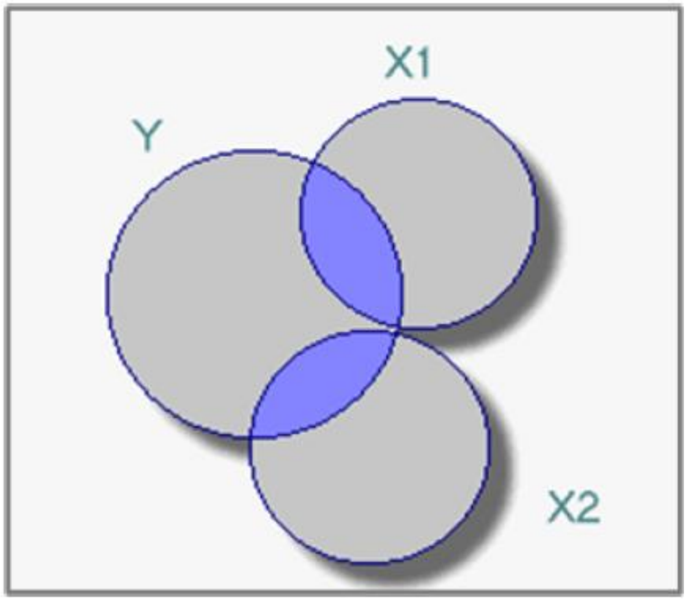
설명변수가 서로 간의 선형결합으로 표현 가능

변수에 대한 가정

2주차 클린업 참고

선형성	해당 가정 위배	설명변수는 확률 변수 X

다중공선성이란?



X1과 X2가 겹치는 부분 없음

▸ 설명 변수 간 독립

X1과 X2가 겹치는 부분 있음

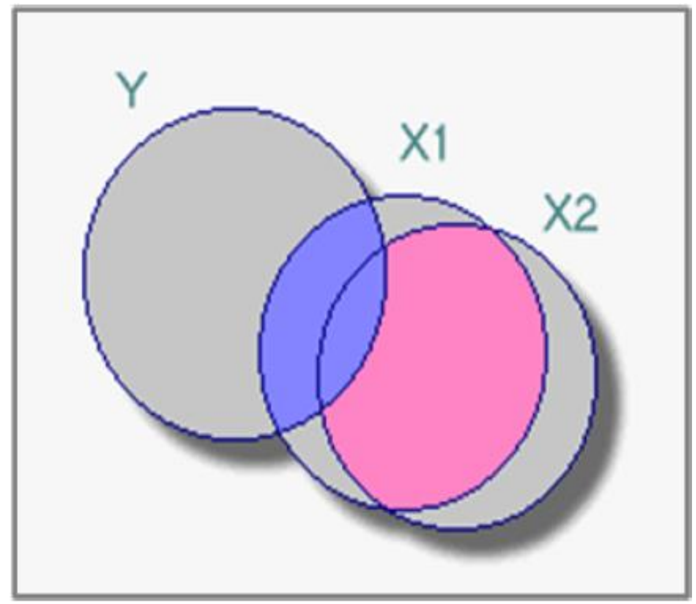
▸ 다중공선성 존재

1 다중공선성

다중공선성이란?

X1과 X2가 겹치는 부분 없음

▸ 설명 변수 간 독립



X1과 X2가 겹치는 부분 있음

▸ **다중공선성** 존재

진단 | ① 직관적인 판단

F-test는 유의하지만 개별 회귀 계수들에 대한 검정에서
귀무가설을 기각하지 못할 경우

상식적으로 유의한 회귀 계수가 유의하지 않다고 나올 경우

추정된 회귀 계수의 부호가 상식과 다를 경우

1 다중공선성

진단 | ① 직관적인 판단

F-test는 유의하지만 개별 회귀 계수들에 대한 검정에서
귀무가설을 기각하지 못할 경우

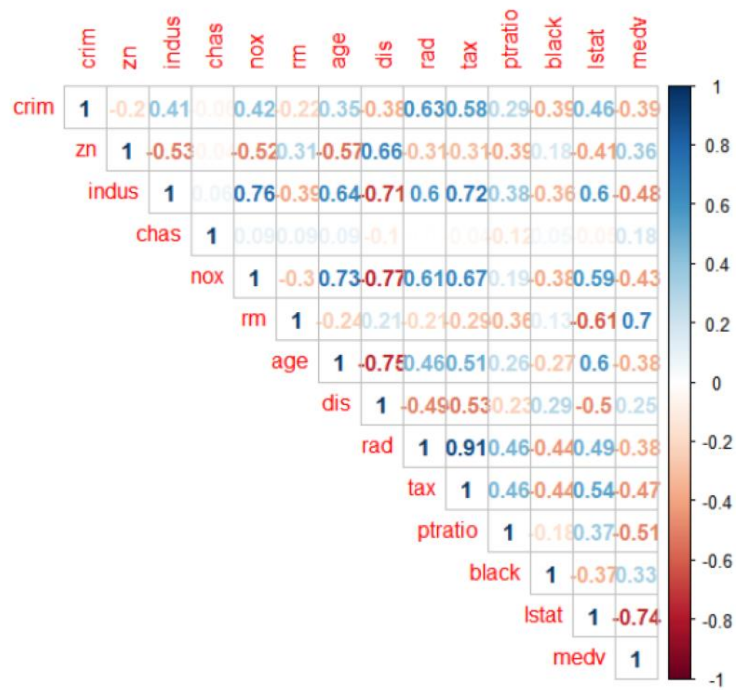
다중공선성이 존재한다고 판단

상식적으로 유의한 회귀 계수가 유의하지 않다고 나올 경우

추정된 회귀 계수의 부호가 상식과 다를 경우

1 다중공선성

진단 | ② 상관계수 플랏



상관계수의 절대값이 0.7 이상 ▸ **다중공선성** 의심

진단 | ③ VIF (분산팽창인자)

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

$\mathrm{R_j^2}$ 란 무엇인가?

다중선형회귀모델을 적합했을 때의 결정계수

$$\mathrm{x_j} = \gamma_1 \mathrm{x_1} + \cdots + \mathrm{r_{j-1}} \mathrm{x}_{j-1} + \mathrm{r_{j+1}} \mathrm{x_{j+1}} + \cdots + \gamma_\mathrm{p} \mathrm{x_p}$$

$\mathrm{R_j^2}$이 큼

x_j가 나머지 변수들의
선형결합으로
충분히 표현 가능

다중공선성 존재

진단 | ③ VIF (분산팽창인자)

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

$\mathrm{R_j^2}$ 란 무엇인가?

다중선형회귀모델을 적합했을 때의 결정계수

$$\mathrm{x_j} = \gamma_1 \mathrm{x_1} + \cdots + \mathrm{r_{j-1}} \mathrm{x}_{j-1} + \mathrm{r_{j+1}} \mathrm{x_{j+1}} + \cdots + \gamma_\mathrm{p} \mathrm{x_p}$$

$\mathrm{R_j^2}$이 큼 → x_j가 나머지 변수들의 선형결합으로 충분히 표현 가능 → 다중공선성 존재

진단 | ③ VIF (분산팽창인자)

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

$\mathrm{R_j^2}$ 란 무엇인가?

다중선형회귀모델을 적합했을 때의 결정계수

$$\mathrm{x_j} = \gamma_1 \mathrm{x_1} + \cdots + \mathrm{r_{j-1}} x_{j-1} + \mathrm{r_{j+1}} \mathrm{x_{j+1}} + \cdots + \gamma_\mathrm{p} \mathrm{x_p}$$

$\mathrm{R_j^2}$이 큼 → x_j가 나머지 변수들의 선형결합으로 충분히 표현 가능 → 다중공선성 존재

1 다중공선성

진단 | ③ VIF (분산팽창인자)

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

MEANING

$\mathrm{VIF} \geq 10$ $(= \mathrm{R_j^2} \geq 0.9)$ 일 경우, 심각한 다중공선성이 있다고 판단

다중공선성이 적을수록 VIF 값은 1에 가까워짐

PCR 의 경우 VIF 들은 모두 1

1 다중공선성

진단 | ③ VIF (분산팽창인자)

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

PCR의 경우, 왜 VIF들은 '1'일까?

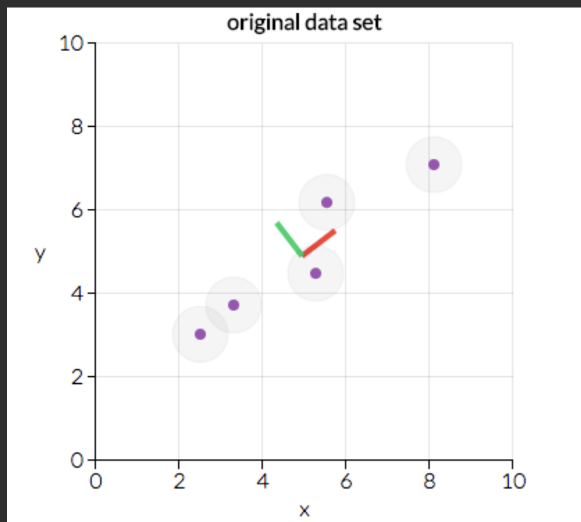
MEANING

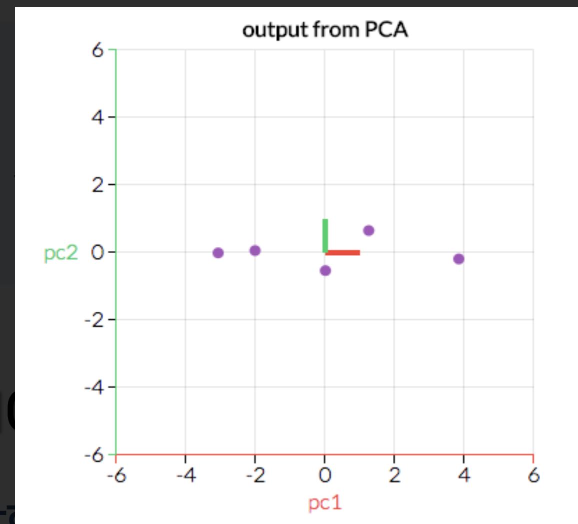
$\mathrm{VIF} \geq 10$ $(= R_j^2 \geq 0.9)$ 일 경우, 심각한 다중공선성이 있다고 판단

다중공선성이 적을수록 VIF 값은 1에 가까워짐

PCR 의 경우 VIF 들은 모두 1

진단 | ③ VIF (분산팽창인자)





PCA : 데이터의 주 성분을 그래프의 주축으로 재설정

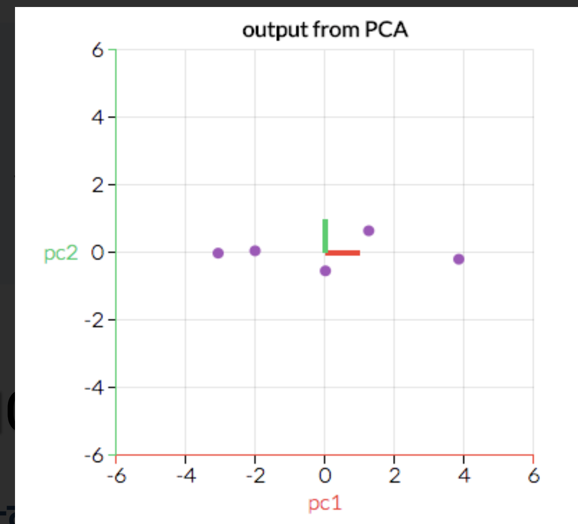
데이터들의 분산이 가장 큰 방향벡터 (빨간선) – PC1

PC1에 수직인 방향벡터(초록선) – PC2

| PCR |

PCA를 기반으로 하는 회귀 분석

데이터의 주성분을 회귀변수로 사용



재정렬된 데이터들은 PC1과 PC2에 대하여 **선형독립**을 이룸

즉, PCR의 VIF는 **1**이 됨

1 다중공선성

문제점 | ① 추정량의 문제

다중선형회귀 : 최소제곱법을 통한 LSE

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

'역행렬' 이 존재하기 위해선?

$X'X$ 가 full rank이어야 함

↓

X가 full-column-rank 행렬이 되어야 함

↓

X의 p개의 변수가 선형 독립이어야 함

1 다중공선성

문제점 | ① 추정량의 문제

다중선형회귀 : 최소제곱법을 통한 LSE

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

'역행렬' 이 존재하기 위해선?

$X'X$ 가 full rank이어야 함

↓

X가 full-column-rank 행렬이 되어야 함

↓

X의 p개의 변수가 선형 독립이어야 함

1 다중공선성

문제점 | ① 추정량의 문제

다중선형회귀 : 최소제곱법을 통한 LSE

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

'역행렬' 이 존재하기 위해선?

$X'X$ 가 full rank이어야 함

↓

X가 full-column-rank 행렬이 되어야 함

↓

X의 p개의 변수가 **선형종속**이라면?

1 다중공선성

문제점 | ① 추정량의 문제

다중선형회귀 : 최소제곱법을 통한 LSE

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

'역행렬' 이 존재하기 위해선?

$X'X$의 **역행렬**은 **존재하지 않음 (다중공선성)**

$X'X$ 가 full rank이어야 함

↓

최소제곱법(OLS method)를 사용할 수 없게 됨

X가 full-column-rank 행렬이 되어야 함

↓

모수의 추정 자체가 어려워짐

X의 p개의 변수가 선형종속이라면?

1 다중공선성

문제점 | ① 추정량의 문제

다중선형회귀 : 최소제곱법을 통한 LSE

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

$X'X$의 역행렬은 존재하지 않음 (다중공선성)

Complete Multicollinearity

최소제곱법(OLS method)를 사용할 수 없게 됨

(완전한 선형종속)

현실에서 완전한 선형종속은 드물며,

근사적으로 선형종속을 이루는 경우가 빈번히 발생

1 다중공선성

문제점 | ① 추정량의 문제

다중선형회귀 : 최소제곱법을 통한 LSE

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

'역행렬' 이 존재하기 위해선?

근사적으로 선형종속이 존재하는 경우

X'X 가 full rank이어야 함

$$\det(X'X) \approx 0$$

X가 full-column-rank 행렬이 되어야 함

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

X의 p개의 변수가 선형종속이라면?

1 다중공선성

문제점 | ① 추정량의 문제

다중선형회귀 : 최소제곱법을 통한 LSE

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

'역행렬' 이 존재하기 위해선?

근사적으로 선형종속이 존재하는 경우

X'X 가 full rank이어야 함

$$\det(X'X) \approx 0$$

X가 full-column-rank 행렬이 되어야 함

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

X의 p개의 변수가 선형종속이라면?

1 다중공선성

문제점 | ① 추정량의 문제

다중선형회귀 : 최소제곱법을 통한 LSE

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

'역행렬' 이 존재하기 위해선?

근사적으로 선형종속이 존재하는 경우

$X'X$ 가 full rank이어야 함

추정량의 분산이 급격하게 커져버려

full column rank 행렬이 되어야 함

계수의 추정이 불안정해짐

X의 p개의 변수가 선형종속이라면?

문제점 | ② 모델의 문제

모델의 검정 결과를 신뢰할 수 없음

다중선형회귀모델이 F-test 통과, R^2값도 괜찮음

But, **유의한 개별 계수**가 **하나도 존재하지 않는 상황 발생**

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

추정량의 분산 커짐 → 검정통계량 감소 (t-test) → 귀무가설을 기각하지 못하게 됨

1 다중공선성

문제점 | ② 모델의 문제

모델의 검정 결과를 신뢰할 수 없음

다중선형회귀모델이 F-test 통과, R^2값도 괜찮음

But, **유의한 개별 계수**가 **하나도 존재하지 않는 상황 발생**

WHY

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

추정량의 분산 커짐 → 검정통계량 감소 (t-test) → 귀무가설을 기각하지 못하게 됨

문제점 | ② 모델의 문제

모델의 해석에 영향을 줌

β_j

다중선형회귀모델

변수 x_j를 제외한 나머지 변수가 고정되어 있을 때,

x_j가 한 단계 증가하면 증가하는 증가량

x_j가 변할 때 선형 종속 관계에 있는 변수들도 변할 수 있음

⇓

"나머지 변수가 고정되어 있을 때"라는 상황이 불가능해짐

1 다중공선성

문제점 | ② 모델의 문제

모델의 해석에 영향을 줌

다중공선성이 존재한다면?

x_j가 변할 때 선형 종속 관계에 있는 변수들도 변할 수 있음

⇓

"나머지 변수가 고정되어 있을 때"라는 상황이 불가능해짐

1 다중공선성

문제점 | ② 모델의 문제

모델의 해석에 영향을 줌

다중공선성이 존재한다면?

x_j가 변할 때 선형 종속 관계에 있는 변수들도 변할 수 있음

⇓

"나머지 변수가 고정되어 있을 때"라는 상황이 **불가능**해짐

1 다중공선성

해결법

다중공선성 해결방법

- 변수 선택
- 정규화
- 차원 축소
- 필터링

1 다중공선성

해결법

다중공선성 해결방법

변수 선택

정규화

차원 축소

필터링

2

변수 선택법

2 변수 선택법

변수 선택법이란?

수많은 변수들 중 적절한 변수 조합을 찾아내는 방법

서로 상관이 있는 독립 변수들을 일부 제거 ⇨ 다중공선성 해결

| 변수 선택법 |

변수가 제거되는 것에

논리성과 정당성을 부여하는 방법

2 변수 선택법

변수 선택법이란?

However,

다중공선성이 완벽하게 제거되는 것은 아님

변수 선택법은 다중공선성만을 해결해주는 방법이 아님

| 변수 선택법 |

변수가 제거되는 것에

논리성과 정당성을 부여하는 방법

2 변수 선택법

변수 선택법이란?

However,

다중공선성이 완벽하게 제거되는 것은 아님

변수 선택법은 다중공선성만을 해결해주는 방법이 아님

| 변수 선택법 |

변수가 제거되는 것에

논리성과 정당성을 부여하는 방법

변수 선택법이란?

However,

Partial F-test 유의성 검정을 통해

변수들을 없애는 방식과의 차이점은?

Partial F-test는 Full Model(FM)과 Reduced Model(RM)이

서로 내포 관계에 있어야 하는 단점!

아래와 같은 경우에서 Partial F-test 불가

$$M_1:\ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad \text{vs} \quad M_2: y = \beta_0' + \beta_3 x_3 + \beta_4 x_4$$

변수가 제거되는 것에

논리성과 정당성을 부여하는 방법

변수 선택법이란?

However,

Partial F-test 유의성 검정을 통해

변수들을 없애는 방식과의 차이점은?

따라서, 일반적인 상황에서 모델의 설명력과 변수의 개수를

모두 고려해 주는 변수 선택 지표를 알아보자!

| 변수 선택법 |

변수가 제거되는 것에

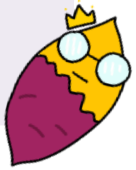
논리성과 정당성을 부여하는 방법

변수 선택 지표

수정결정계수 R^2_{adj}

설명력을 담당하는 결정계수와 변수 개수 패널티가

수정결정계수 식에 포함



회귀분석 클린업 1주차 다중선형**회귀**의
적합성 검정파트로 **회귀**해보자

수정결정계수 식

$$1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

변수 선택 지표

AIC Akaike Information Criterion

일반적인 $\text{AIC} = -2\log(\text{Likelihood}) + 2k$

정규분포 가정 하에서의 $\text{AIC} = n\log(2\pi\hat{\sigma}^2) + \frac{SSE}{\hat{\sigma}^2} + 2k$

$\hat{\sigma}^2$: σ^2의 MLE

Likelihood : 값이 커질 수록 , 모델이 데이터를 잘 설명한다는 의미

K : 모수의 개수로 변수 개수에 따른 패널티를 부과한 것

Mallow's Cp

정규성과 선형성 가정 하에 AIC와 동일한 지표

변수 선택 지표

AIC Akaike Information Criterion

일반적인 $\text{AIC} = -2\log(\text{Likelihood}) + 2k$

정규분포 가정 하에서의 $\text{AIC} = n\log(2\pi\hat{\sigma}^2) + \frac{SSE}{\hat{\sigma}^2} + 2k$

$\hat{\sigma}^2$: σ^2의 MLE

Likelihood : 값이 커질 수록 , 모델이 데이터를 잘 설명한다는 의미

K : 모수의 개수로 변수 개수에 따른 패널티를 부과한 것

Mallow's Cp

정규성과 선형성 가정 하에 AIC와 동일한 지표

변수 선택 지표

BIC Bayesian Information Criterion

일반적인 $\mathrm{BIC} = -2\log(\mathrm{Likelihood}) + \mathrm{klog}(n)$

정규분포 가정 하에서의 $\mathrm{BIC} = \mathrm{nlog}(2\pi\hat{\sigma}^2) + \frac{SSE}{\hat{\sigma}^2} + \mathrm{klog}(n)$

AIC와 유사함

$\hat{\sigma}^2$: σ^2의 MLE

Likelihood : 값이 커질 수록 , 모델이 데이터를 잘 설명한다는 의미

K : 모수의 개수로 변수 개수에 따른 패널티를 부과한 것

데이터의 개수를 모수의 개수에 곱함으로써

AIC보다 더 큰 패널티 부과

변수 선택법

Best Subset Selection

가능한 모든 변수들의 조합을 다 고려하는 방법

변수의 개수가 p개일 때, 2^p 개의 모형을 모두 적합하고 비교

Positive

가능한 모든 경우의 수를 고려

선택된 best model을 신뢰할 수 있음

Negative

계산 비용이 많이 소모

변수의 개수가 40개를 초과하면 불가능

변수 선택법

Best Subset Selection

1. $M_1, \dots, M_P$ 개의 모형을 적합

   M_k의 모형은 k개의 변수를 포함하는 모형 중 training error를 작게 하는 모형

2. P개의 모형 중 AIC / BIC가 가장 작은 모형을 선택

Positive

가능한 모든 경우의 수를 고려

선택된 best model을 신뢰할 수 있음

Negative

계산 비용이 많이 소모

변수의 개수가 40개를 초과하면 불가능

변수 선택법

Best Subset Selection

1. M_1 , … , M_P 개의 모형을 적합

   M_k의 모형은 k개의 변수를 포함하는 모형 중 training error를 작게 하는 모형

2. P개의 모형 중 AIC / BIC가 가장 작은 모형을 선택

Positive

가능한 **모든 경우의 수를 고려**

선택된 best model을 신뢰할 수 있음

Negative

계산 비용이 많이 소모

변수의 개수가 40개를 초과하면 불가능

변수 선택법

전진선택법 Forward Selection

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

Positive

계산이 매우 빠름

N < P 인 경우에도 사용 가능

Negative

가능한 모든 변수 조합을 고려하지 않음

선택된 모형이 최적의 모형 X

변수 선택법

전진선택법 Forward Selection

1. Null Model에서 변수를 하나 추가하여 1개의 변수를 포함하고 있는 모형들 중 MSE를 가장 작게 만드는 모형 M_1를 선택

Positive

계산이 매우 빠름

N < P 인 경우에도 사용 가능

Negative

가능한 모든 변수 조합을 고려하지 않음

선택된 모형이 최적의 모형 X

변수 선택법

전진선택법 Forward Selection

2. 변수를 하나씩 추가하면서 MSE를 비교하여 총 p개의 모형을 적합

3. M_1 , … , M_P 개의 모델 중 AIC / BIC가 가장 작은 모형을 선택

Positive

계산이 매우 빠름

N < P 인 경우에도 사용 가능

Negative

가능한 모든 변수 조합을 고려하지 않음

선택된 모형이 최적의 모형 X

변수 선택법

전진선택법 Forward Selection

2. 변수를 하나씩 추가하면서 MSE를 비교하여 총 p개의 모형을 적합

3. M_1 , … , M_P 개의 모델 중 AIC / BIC가 가장 작은 모형을 선택

Positive

계산이 매우 빠름

N < P 인 경우에도 사용 가능

Negative

가능한 모든 변수 조합을 고려하지 않음

선택된 모형이 최적의 모형 X

변수 선택법

후진제거법 Backward Elimination

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

변수를 하나씩 제거하는 방법

Positive

Best subset selection에 비해

계산이 매우 빠름

Negative

선택된 모형이 최적의 모형 X

N < P 인 경우에는 사용 불가능

$(X'X)^{-1}$가 존재하지 않기 때문!

변수 선택법

후진제거법 Backward Elimination

1. Full Model에서 변수를 하나 제거하여 p-1개의 변수를 포함하고 있는 모형들 중 MSE를 가장 작게 만드는 모형 M_{p-1}를 선택

Positive

Best subset selection에 비해 계산이 매우 빠름

Negative

선택된 모형이 최적의 모형 X

N < P 인 경우에는 사용 불가능

$(X'X)^{-1}$가 존재하지 않기 때문!

변수 선택법

후진제거법 Backward Elimination

2. 1개씩 제거 하여 총 p개의 모형을 적합

3. M_1 , … , M_P 개의 모델 중 AIC / BIC가 가장 작은 모형을 선택

Positive

Best subset selection에 비해

계산이 매우 빠름

Negative

선택된 모형이 최적의 모형 X

N < P 인 경우에는 사용 불가능

$(X'X)^{-1}$가 존재하지 않기 때문!

변수 선택법

후진제거법 Backward Elimination

2. 1개씩 제거 하여 총 p개의 모형을 적합

3. $M_1, \ldots, M_P$ 개의 모델 중 AIC / BIC가 가장 작은 모형을 선택

Positive

Best subset selection에 비해

계산이 매우 빠름

Negative

선택된 모형이 최적의 모형 X

N < P 인 경우에는 사용 불가능

$(X'X)^{-1}$가 존재하지 않기 때문!

변수 선택법

단계적선택법 Stepwise Selection

전진선택법과 후진제거법 과정을 섞은 방법

Positive

Best subset selection에 비해

계산이 매우 빠름

Negative

선택된 모형이 최적의 모형 X

N < P 인 경우에는 제한 있음

$(X'X)^{-1}$가 존재하지 않기 때문!

변수 선택법

단계적선택법 Stepwise Selection

1. Null model 또는 Full model에서 시작
2. 다른 변수 선택법들을 혼합하여 변수를 제거/추가하며 모델을 평가
3. AIC / BIC가 가장 작은 모형을 선택

Positive

Best subset selection에 비해

계산이 매우 빠름

Negative

선택된 모형이 최적의 모형 X

N < P 인 경우에는 제한 있음

$(X'X)^{-1}$가 존재하지 않기 때문!

변수 선택법

단계적선택법 Stepwise Selection

1. Null model 또는 Full model에서 시작
2. 다른 변수 선택법들을 혼합하여 변수를 제거/추가하며 모델을 평가
3. AIC / BIC가 가장 작은 모형을 선택

Positive

Best subset selection에 비해
계산이 매우 빠름

Negative

선택된 모형이 최적의 모형 X
N < P 인 경우에는 제한 있음
$(X'X)^{-1}$가 존재하지 않기 때문!

변수 선택법

단계적선택법 Stepwise Selection

1. Null model 또는 Full model에서 시작
2. 다른 변수 선택법들을 혼합하여 변수를 제거/추가하며 모델을 평가
3. AIC / BIC가 가장 작은 모형을 선택

Positive

Best subset selection에 비해

계산이 매우 빠름

Negative

선택된 모형이 최적의 모형 X

N < P 인 경우에는 제한 있음

$(X'X)^{-1}$가 존재하지 않기 때문!

변수 선택법의 한계

전진선택법과 후진제거법의 결과가 상이할 수 있음

위의 두가지 방법은 모든 변수 조합을 고려하는 것이 아니기 때문에

기계적으로 변수를 제거하는 행위는 위험

정규화 방법 추천

3

정규화

정규화

Regularization

회귀 계수가 가질 수 있는 값에 제약 조건을 부여함으로써

계수들을 작게 만들거나 0으로 만드는 방법

Bias-vaiance trade off

다중공선성이 존재하는 경우

회귀계수들의 OLS 추정량

정규화

Effect

다중공선성은 OLS 추정량의 분산을 크게 증가시킴

정규화는 OLS 추정량의 불편성 포기 ▶ 분산을 줄이는 효과가 있음

Bias-vaiance trade off

다중공선성이 존재하는 경우

회귀계수들의 OLS 추정량

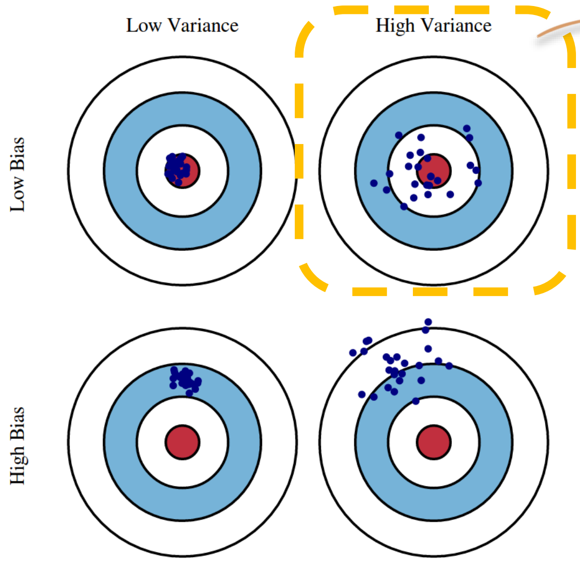
정규화

Effect

다중공선성은 OLS 추정량의 분산을 크게 증가시킴

정규화는 OLS 추정량의 불편성 포기 ▶ 분산을 줄이는 효과가 있음

Bias-vaiance trade off



다중공선성이 존재하는 경우

회귀계수들의 OLS 추정량

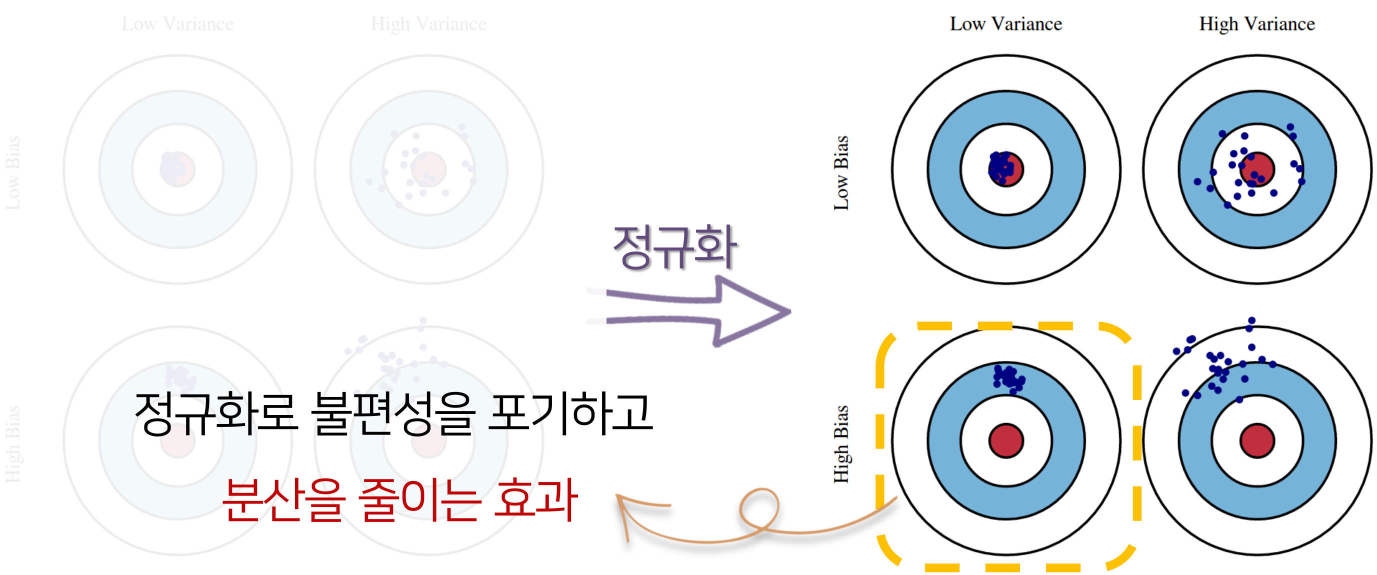
정규화

Effect

다중공선성은 OLS 추정량의 분산을 크게 증가시킴

정규화는 OLS 추정량의 불편성 포기 ▶ 분산을 줄이는 효과가 있음

Bias-vaiance trade off



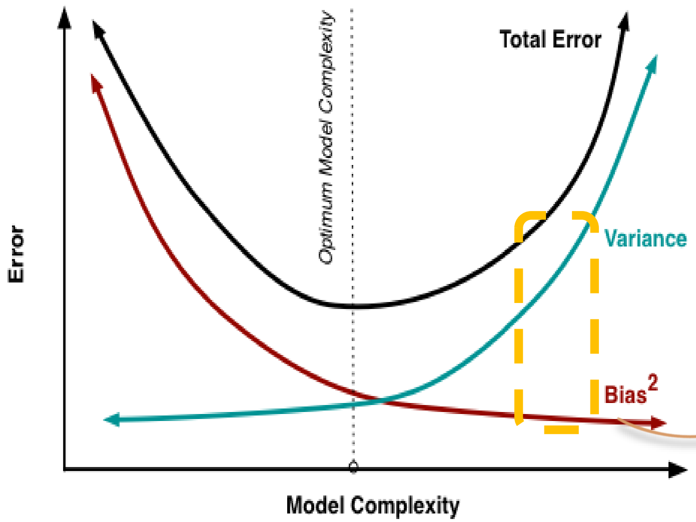
정규화

Effect

다중공선성은 OLS 추정량의 분산을 크게 증가시킴

정규화는 OLS 추정량의 불편성 포기 ▶ 분산을 줄이는 효과가 있음

Bias-vaiance trade off



다중 공선성이 존재하는 경우

모델의 분산과 편향

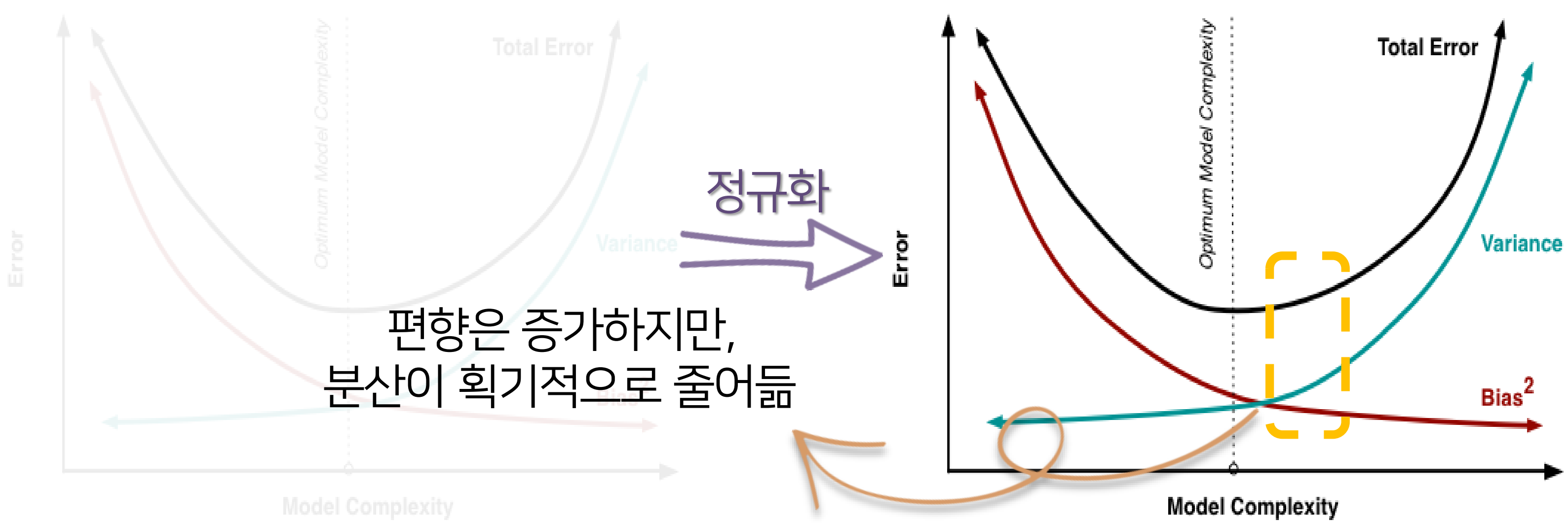
정규화

Effect

다중공선성은 OLS 추정량의 분산을 크게 증가시킴

정규화는 OLS 추정량의 불편성 포기 ▶ 분산을 줄이는 효과가 있음

Bias-vaiance trade off



Ridge Regression

Ridge Regression

SSE를 최소화하면서 회귀계수에 제약 조건을 거는 방법

(L2 Regularization)

목적함수

$$\operatorname{argmin} \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_j\right)^2 \text{ subject to } ||\beta||_2^2 \le s$$

$$\operatorname{argmin} \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_j\right)^2 + \lambda||\beta||_2^2$$

목적함수를 최소화함으로써 Ridge Estimator 추정 가능

이차식 형태이므로 미분을 통해 추정량 계산

Ridge Lasso Elastic-Net Fused Lasso

Ridge Regression

왜 L2 Regularization인가?

제약 조건식이 L2-norm 형태

$$\|x\|_2 = (x_1^2 + \cdots + x_n^2)^{1/2}$$

목적함수

$$\operatorname{argmin} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j\right)^2 \text{ subject to } ||\beta||_2^2 \le s$$

$$\operatorname{argmin} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j\right)^2 + \lambda||\beta||_2^2$$

목적함수를 최소화함으로써 Ridge Estimator 추정 가능

이차식 형태이므로 미분을 통해 추정량 계산

Ridge Regression

왜 L2 Regularization인가?

제약 조건식이 L2-norm 형태

$$\|x\|_2 = (x_1^2 + \cdots + x_n^2)^{1/2}$$

목적함수

$$\operatorname{argmin} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j\right)^2 \text{ subject to } \|\beta\|_2^2 \leq s$$

$$\operatorname{argmin} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j\right)^2 + \lambda \|\beta\|_2^2$$

목적함수를 최소화함으로써 Ridge Estimator 추정 가능

이차식 형태이므로 미분을 통해 추정량 계산

Ridge Regression

왜 L2 Regularization인가?

제약 조건식이 L2-norm 형태

$$\|x\|_2 = (x_1^2 + \cdots + x_n^2)^{1/2}$$

목적함수

$$\operatorname{argmin} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j\right)^2 \text{ subject to } \|\beta\|_2^2 \leq s$$

$$\operatorname{argmin} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j\right)^2 + \lambda \|\beta\|_2^2$$

목적함수를 최소화함으로써 Ridge Estimator 추정 가능

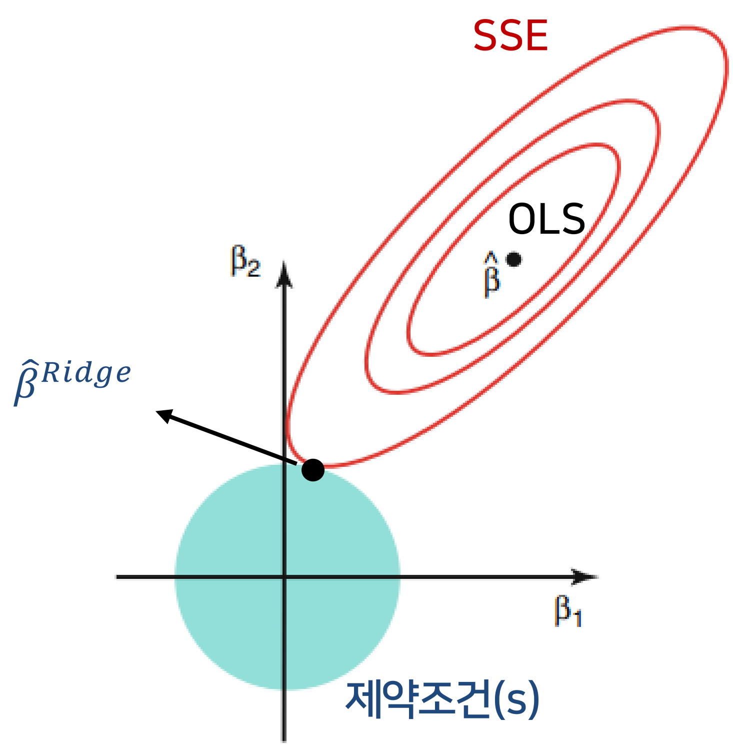
⚠ 설명 변수들은 **표준화된 상태**여야 함

이차식 형태이므로 미분을 통해 추정량 계산

Ridge Lasso Elastic-Net Fused Lasso

목적함수에 대한 이해 ①

$$\text{argmin} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j\right)^2 \text{ subject to } ||\beta||_2^2 \leq s$$



기본 요소

파란 원 : 제약 조건 in Ridge

붉은 타원 : 추정량의 SSE

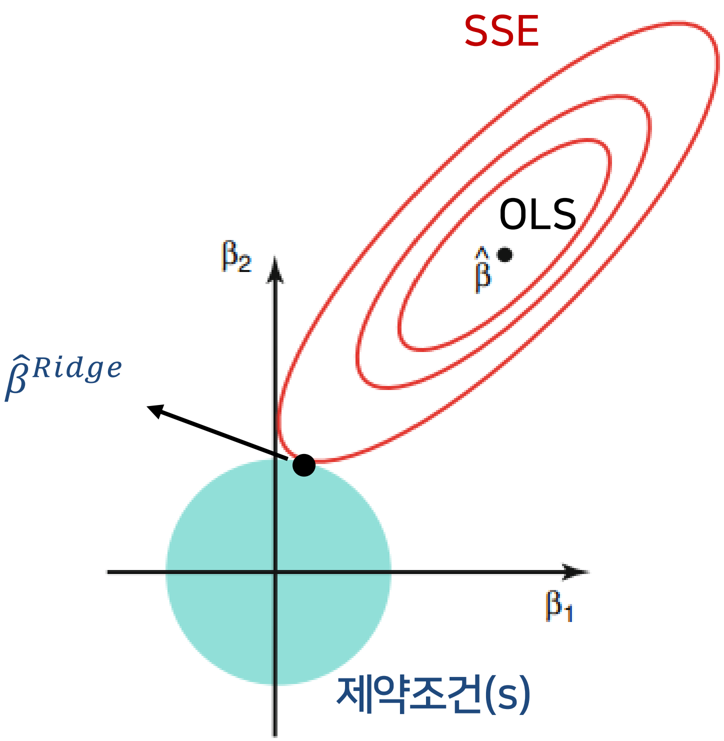
내부 점 : OLS 추정량

회귀계수의 최소화

1. 회귀계수 $\hat{\beta}$는 반드시 원 내부에 존재
2. SSE를 최소화해야 함

목적함수에 대한 이해 ①

$$\text{argmin} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j\right)^2 \text{ subject to } \|\beta\|_2^2 \leq s$$



기본 요소

파란 원 : 제약 조건 in Ridge

붉은 타원 : 추정량의 SSE

타원과 원의 접점이 Ridge Estimator!

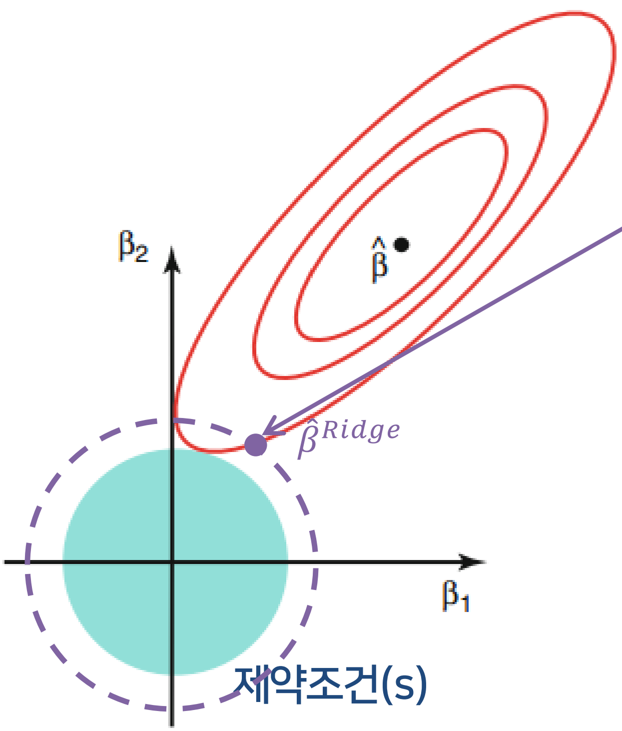
회귀계수의 최소화

1. 회귀계수 $\hat{\beta}$는 반드시 원 내부에 존재

2. SSE를 최소화해야 함

Ridge Lasso Elastic-Net Fused Lasso

목적함수에 대한 이해 ①



제약 조건이 완화될 경우

s가 증가 > 원의 넓이 증가

원이 타원을 밀어내며 추정량이 0에서 멀어짐

회귀 계수를 작게 만들 수 없음

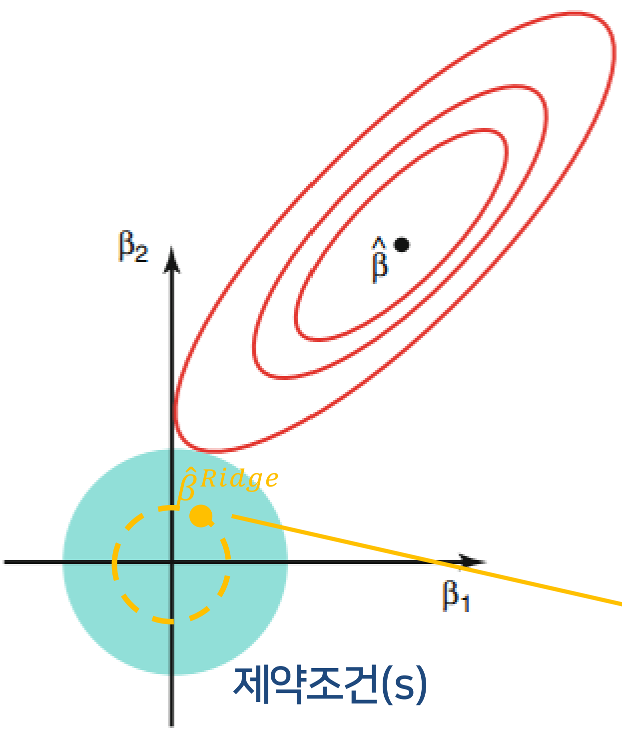
제약 조건이 강화될 경우

s가 감소 > 원의 넓이 감소

추정량이 0으로 수렴함 (0은 될 수 없음)

회귀 계수를 작게 만들 수 있음

목적함수에 대한 이해 ①



제약 조건이 완화될 경우

s가 증가 > 원의 넓이 증가

원이 타원을 밀어내며 추정량이 0에서 멀어짐

회귀 계수를 작게 만들 수 없음

제약 조건이 강화될 경우

s가 감소 > 원의 넓이 감소

추정량이 0으로 수렴함 (0은 될 수 없음)

회귀 계수를 작게 만들 수 있음

목적함수에 대한 이해 ②

라그랑지안 승수법을 이용해 나타낸 함수식

$$\operatorname{argmin} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j\right)^2 + \lambda \|\beta\|_2^2$$

오차제곱합을 최소화 & Penalty term을 통해

회귀계수의 크기 조정

👑Tuning Parameter👑

음수가 아닌 튜닝파라미터

최적의 모델을 찾는 과정에서 직접 CV를 통해 조정해주는 모수

제약조건의 크기를 결정 (s와는 반대 관계)

Ridge Lasso Elastic-Net Fused Lasso

목적함수에 대한 이해 ②

라그랑지안 승수법을 이용해 나타낸 함수식

$$\text{argmin} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j\right)^2 + \lambda \|\beta\|_2^2$$

오차제곱합을 최소화 & Penalty term을 통해

회귀계수의 크기 조정

Tuning Parameter

음수가 아닌 튜닝파라미터

최적의 모델을 찾는 과정에서 직접 CV를 통해 조정해주는 모수

제약조건의 크기를 결정 (s와는 반대 관계)

목적함수에 대한 이해 ②

Ridge Lasso Elastic-Net Fused Lasso

λ의 값에 따른 회귀계수의 변화

라그랑지안 승수법을 이용해 나타낸 함수식

$$\text{argmin} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j \right)^2 + \lambda \|\beta\|_2^2$$

오차제곱합을 최소화 & Penalty term을 통해 회귀계수의 크기 조정

λ가 커지는 경우

λ의 영향력이 증가하므로,

전체 식을 최소화하기 위해

$\|\beta\|_2^2$가 작아져야 함

∴ 개별 회귀 계수들은 감소

λ가 작아지는 경우

λ의 영향력이 감소하므로,

상대적으로 $\|\beta\|_2^2$의 영향력 증가

∴ 개별 회귀 계수들은 증가

Tuning Parameter

음수가 아닌 튜닝파라미터

최적화 과정에서 직접 CV를 통해 조정해주는 모수

제약조건의 크기를 결정 (s와는 반대 관계)

Ridge Lasso Elastic-Net Fused Lasso

목적함수에 대한 이해 ②

λ의 값에 따른 회귀계수의 변화

라그랑지안 승수법을 이용해 나타낸 함수식

$$\operatorname{argmin} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j\right)^2 + \lambda \|\beta\|_2^2$$

오차제곱합을 최소화 & Penalty term을 통해 회귀계수의 크기 조정

λ가 커지는 경우

λ의 영향력이 증가하므로,

전체 식을 최소화하기 위해

$\|\beta\|_2^2$가 작아져야 함

∴ 개별 회귀 계수들은 감소

λ가 작아지는 경우

$\lambda = 0$ 이 된다면

Panelty term 없어짐

∴ OLS 추정량과 동일

Tuning Parameter

음수가 아닌 튜닝파라미터

…정에서 직접 CV를 통해 조정해주는 모수

제약조건의 크기를 결정 (s와는 반대 관계)

Ridge Lasso Elastic-Net Fused Lasso

목적함수에 대한 이해 ③

$$\underbrace{\operatorname{argmin}\,(\mathbf{y} - \boldsymbol{X\beta})'(\mathbf{y} - \boldsymbol{X\beta})}_{\text{SSE}} + \underbrace{\boldsymbol{\lambda\beta'\beta}}_{\text{Penalty}}$$

다중공선성이 존재하는 경우,

$\boldsymbol{X'X}$ 열들 사이에 강한 선형 관계가 존재

⇓

회귀 계수 추정량

$$\hat{\beta}^{ridge} = (\boldsymbol{X}'\mathbf{X} + \boldsymbol{\lambda}\mathbf{I})^{-1}\boldsymbol{X}'\mathbf{y}$$

$\boldsymbol{\lambda}\mathbf{I}$ 를 더해주어

$\boldsymbol{X}'\mathbf{X} + \boldsymbol{\lambda}\mathbf{I}$ 를 Full-rank Matrix로 만들어 줌

목적함수에 대한 이해 ③

$$\underbrace{\text{argmin}\,(\mathbf{y} - \boldsymbol{X\beta})'(\mathbf{y} - \boldsymbol{X\beta})}_{\text{SSE}} + \underbrace{\boldsymbol{\lambda\beta'\beta}}_{\text{Penalty}}$$

다중공선성이 존재하는 경우,

$\boldsymbol{X'X}$ 열들 사이에 강한 선형 관계가 존재

$\Downarrow$

회귀 계수 추정량

$$\hat{\beta}^{ridge} = (\boldsymbol{X'}\mathbf{X} + \boldsymbol{\lambda}\mathbf{I})^{-1}\boldsymbol{X'}\mathbf{y}$$

$\boldsymbol{\lambda}\mathbf{I}$ 를 더해주어

$\boldsymbol{X'}\mathbf{X} + \boldsymbol{\lambda}\mathbf{I}$ 를 Full-rank Matrix로 만들어 줌

Ridge Lasso Elastic-Net Fused Lasso

특징

Scaling

계수는 변수 단위에 가장 큰 영향 받음

단위의 영향을 제거하고 순수 영향만 표현

주로 standard scaling 사용

계산 비용 절약

Penalty Term 덕분에 미분이 가능

λ를 바꾸며 미분과 함께 행렬 연산

예측 성능

모델에 다중공선성이 존재할 경우,

높은 예측 성능을 보임

변수 선택

영향력을 줄일 뿐 변수는 잔존

다중공선성을 일으키는 변수는 제거 불가

Ridge를 통한 해석력 증가는 어려움

특징

Ridge Lasso Elastic-Net Fused Lasso

Scaling

계수는 변수 단위에 가장 큰 영향 받음

단위의 영향을 제거하고 순수 영향만 표현

주로 standard scaling 사용

계산 비용 절약

Penalty Term 덕분에 미분이 가능

λ를 바꾸며 미분과 함께 행렬 연산

예측 성능

모델에 다중공선성이 존재할 경우,

높은 예측 성능을 보임

변수 선택

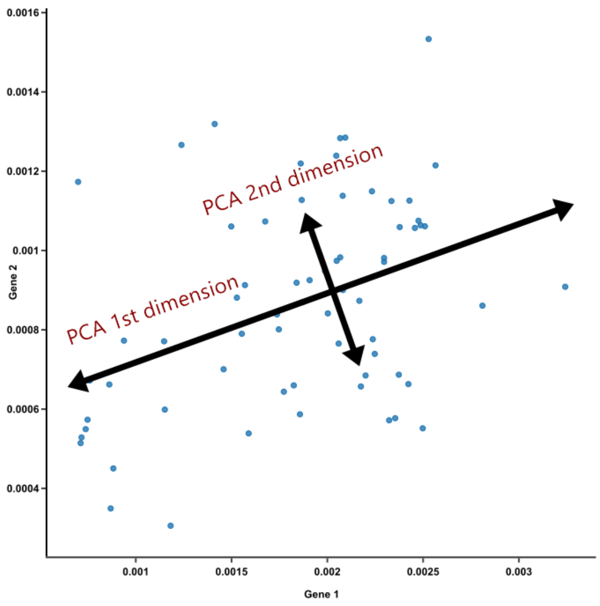
영향력을 줄일 뿐 변수는 잔존

다중공선성을 일으키는 변수는 제거 불가

Ridge를 통한 해석력 증가는 어려움

PCR과 Ridge의 비교

Ridge regression = PCR의 smoothing Version



PCA를 회귀분석에 적용한 방법

선형대수학 3주차 클린업 참고

X의 분산-공분산 행렬의

고유벡터를 축으로 하는 공간에

데이터를 투영하여 차원을 감소시킴

PCR과 Ridge의 비교

특이값 분해(SVD) $\mathbf{X} = \boldsymbol{USV'}$ 형태로 X행렬을 분해

PCR

$$\hat{\boldsymbol{y}}_{\boldsymbol{PCR}} = \boldsymbol{X}_{\boldsymbol{PCA}}\hat{\boldsymbol{\beta}}_{\boldsymbol{PCR}} = \mathbf{U} \cdot \mathrm{diag}\left(1_1, \dots, 1_k, 0_{k+1}, \dots 0_p\right) \cdot \boldsymbol{U}'\mathbf{y}, \quad \boldsymbol{X}_{\boldsymbol{PCA}} = \mathbf{US}$$

대각행렬에서 차이가 존재

Ridge

$$\hat{\boldsymbol{y}}_{\boldsymbol{ridge}} = \boldsymbol{X}_{\boldsymbol{ridge}}\hat{\boldsymbol{\beta}}_{\boldsymbol{ridge}} = \mathbf{X}\left(\boldsymbol{X}'\mathbf{X} + \boldsymbol{\lambda I}\right)^{-1}\boldsymbol{X}'\mathbf{y} = \mathbf{U} \cdot \mathrm{diag}\left(\frac{s_i^2}{s_i^2 + \lambda}\right) \cdot \boldsymbol{U}'\mathbf{y}$$

PCR과 Ridge의 비교

특이값 분해(SVD) $\mathbf{X} = \boldsymbol{USV'}$ 형태로 X행렬을 분해

PCR

$$\hat{\boldsymbol{y}}_{\boldsymbol{PCR}} = \boldsymbol{X}_{\boldsymbol{PCA}}\hat{\boldsymbol{\beta}}_{\boldsymbol{PCR}} = \mathbf{U}\cdot \mathrm{diag}\left(1_1, \dots, 1_k, 0_{k+1}, \dots 0_p\right)\cdot \boldsymbol{U'}\mathbf{y}, \quad \boldsymbol{X}_{\boldsymbol{PCA}} = \mathbf{US}$$

임의로 선택한 PC의 개수

Ridge

$$\hat{\boldsymbol{y}}_{\boldsymbol{ridge}} = \boldsymbol{X}_{\boldsymbol{ridge}}\hat{\boldsymbol{\beta}}_{\boldsymbol{ridge}} = \mathbf{X}\,(\boldsymbol{X'}\mathbf{X} + \boldsymbol{\lambda I})^{-1}\boldsymbol{X'}\mathbf{y} = \mathbf{U}\cdot \mathrm{diag}\left(\frac{s_i^2}{s_i^2 + \lambda}\right)\cdot \boldsymbol{U'}\mathbf{y}$$

X의 분산-공분산 행렬의 i 번째 고유 벡터가

담고 있는 정보의 크기

PCR과 Ridge의 비교

특이값 분해(SVD) $\mathbf{X} = \boldsymbol{USV'}$ 형태로 X행렬을 분해

PCR

s_i^2 : X의 분산-공분산 행렬 고유벡터가 담고 있는 정보의 크기

Ridge

정보량이 많음 > s_i^2의 크기 λ 의 영향력 > 1에 가까움

정보량이 적음 > s_i^2의 크기 λ 의 영향력 > 0에 가까움

$$\mathrm{diag}\left(\frac{s_i^2}{s_i^2 + \lambda}\right)$$

PCR과 Ridge의 비교

특이값 분해(SVD) $\mathbf{X} = \boldsymbol{USV'}$ 형태로 X행렬을 분해

PCR

정보량이 많음 $\lambda = 0$

$\mathrm{diag}(1_1, \ldots, 1_k, 0_{k+1}, \ldots 0_p)$

정보량이 적음 $\lambda = \infty$

극단적으로 λ를 선택

Ridge

정보량이 많음 > s_i^2의 크기 ↑ λ 의 영향력 ↓ > 1에 가까움

정보량이 적음 > s_i^2의 크기 ↓ λ 의 영향력 ↑ > 0에 가까움

$$\mathrm{diag}\left(\frac{s_i^2}{s_i^2 + \lambda}\right)$$

PCR과 Ridge의 비교

특이값 분해(SVD) $\mathbf{X} = \boldsymbol{USV'}$ 형태로 X행렬을 분해

PCR

$$\lambda \to 0 \qquad \mathrm{diag}\left(1_1, \ldots, 1_k, 0_{k+1}, \ldots 0_p\right) \qquad \lambda \to \infty$$

극단적으로 λ를 선택

Ridge

$$\mathrm{diag}\left(\frac{s_i^2}{s_i^2 + \lambda}\right)$$

∴ Ridge regression이 PCR의 smoothing version

Ridge **Lasso** Elastic-Net Fused Lasso

Lasso Regression

Lasso Regression

SSE를 최소화하면서 회귀계수 β에 제약 조건을 거는 방법

(L1 Regularization)

제약 조건식이 L1-norm 형태

목적함수

$$\operatorname{argmin} \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j\right)^2 \text{ subject to } \|\beta\|_1 \le s$$

$$\operatorname{argmin} \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j\right)^2 + \lambda\|\beta\|_1$$

식을 최소화하여 회귀계수의 Lasso estimator를 얻을 수 있음

Lasso Regression

Lasso Regression

SSE를 최소화하면서 회귀계수 β에 제약 조건을 거는 방법

(L1 Regularization)

제약 조건식이 L1-norm 형태

목적함수

$$\operatorname{argmin} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j \right)^2 \text{ subject to } \|\beta\|_1 \leq s$$

$$\operatorname{argmin} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j \right)^2 + \lambda \|\beta\|_1$$

식을 최소화하여 회귀계수의 Lasso estimator를 얻을 수 있음

Lasso Regression

Lasso Regression

SSE를 최소화하면서 회귀계수 β에 제약 조건을 거는 방법

(L1 Regularization)

제약 조건식이 L1-norm 형태

목적함수

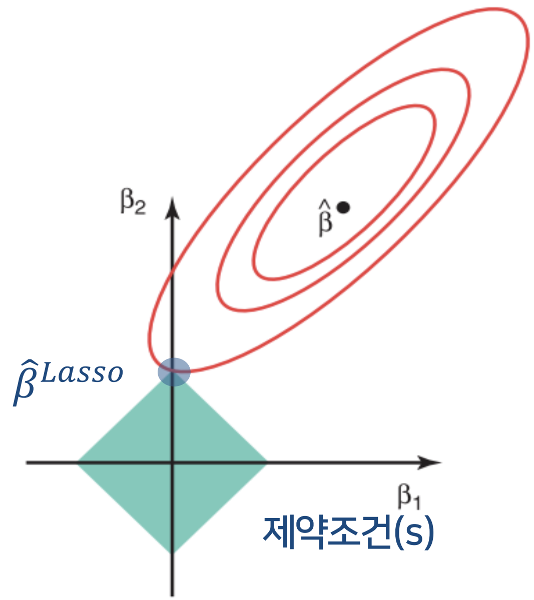
Lasso ▶ 미분 불가능

수치적인 방법을 통해 회귀 계수 Lasso estimator를 구해야 함 !

(설명 변수들은 표준화된 상태)

목적함수에 대한 이해 ①

$$\operatorname{argmin} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_j\right)^2 \text{ subject to } \|\beta\|_1 \le s$$

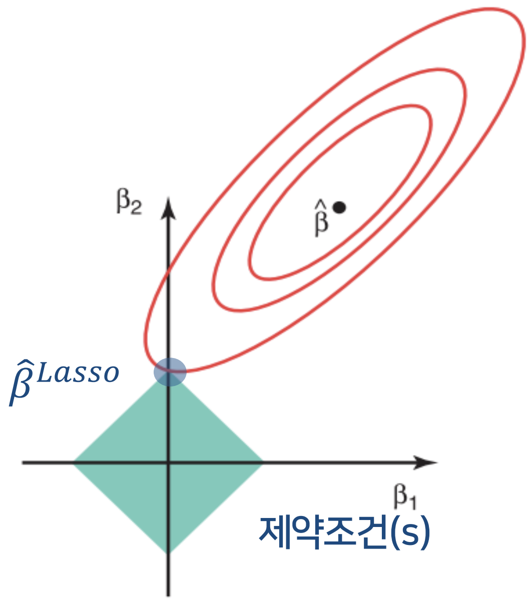


Ridge와 달리 제약 조건의 형태가 마름모꼴

제약조건을 만족시키는 동시에 SSE 최소화

목적함수에 대한 이해 ①

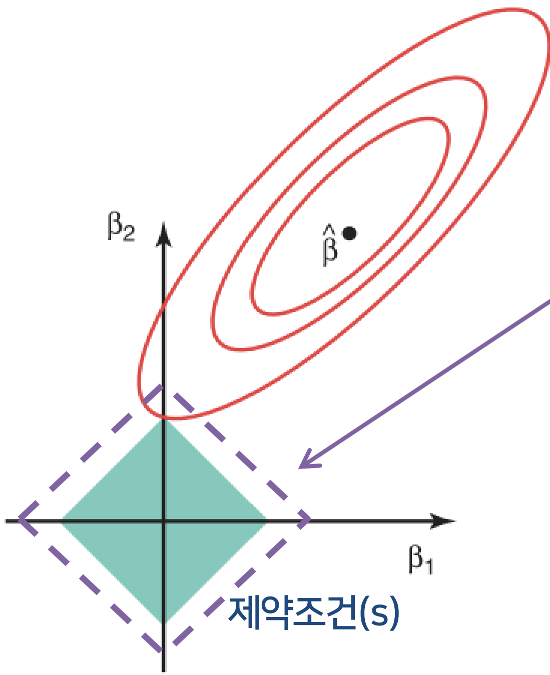
$$\text{argmin} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j\right)^2 \text{ subject to } ||\beta||_1 \leq s$$



타원과 마름모의 접점

= 회귀계수의 Lasso estimator

목적함수에 대한 이해 ①



제약 조건이 완화될 경우

마름모의 넓이 s가 증가

마름모가 타원을 밀어내며 추정량이 0에서 멀어짐

회귀 계수를 작게 만들 수 없음

제약 조건이 강화될 경우

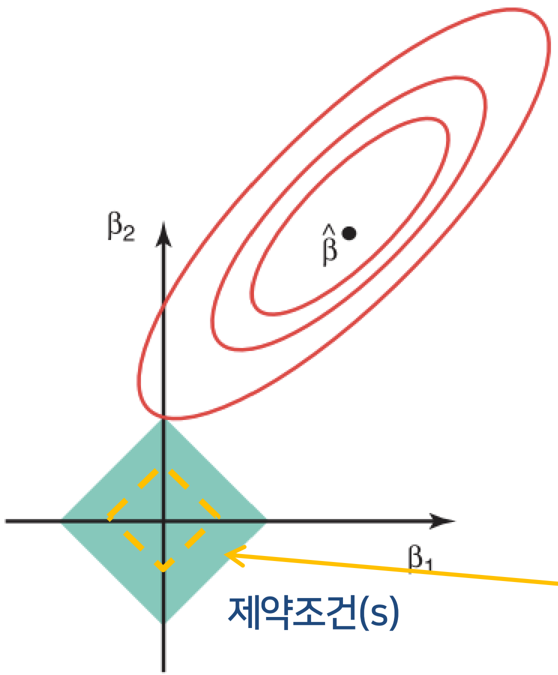
마름모의 넓이 s가 감소

추정량이 0으로 수렴함 (0이 될 수 있음)

회귀 계수를 작게 만들 수 있음

Ridge Lasso Elastic-Net Fused Lasso

목적함수에 대한 이해 ①



제약 조건이 완화될 경우

마름모의 넓이 s가 증가

마름모가 타원을 밀어내며 추정량이 0에서 멀어짐

회귀 계수를 작게 만들 수 없음

제약 조건이 강화될 경우

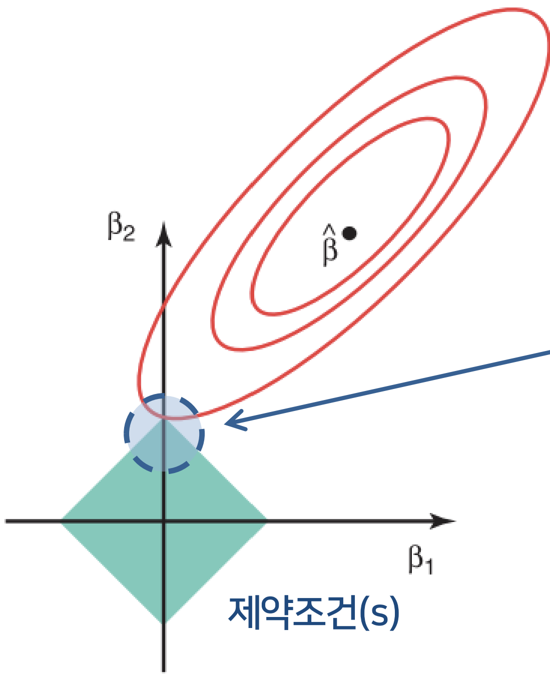
마름모의 넓이 s가 감소

추정량이 0으로 수렴함 (0이 될 수 있음)

회귀 계수를 작게 만들 수 있음

Ridge Lasso Elastic-Net Fused Lasso

목적함수에 대한 이해 ①



제약 조건이 완화될 경우

마름모의 넓이 s가 증가

Ridge와 달리

마름모꼴의 제약 조건으로 인해

일부 회귀 계수 = 0 인 가능성 존재

제약 조건이 강화될 경우

마름모의 넓이 s가 감소

추정량이 0으로 수렴함 (0이 될 수 있음)

회귀 계수를 작게 만들 수 있음

목적함수에 대한 이해 ②

라그랑지안 승수법을 이용해 나타낸 함수식

$$\underbrace{\operatorname{argmin} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j\right)^2}_{\text{SSE}} + \underbrace{\lambda \|\beta\|_1}_{\text{Penalty}}$$

개별 회귀 계수가 너무 많아지는 것을 조정

(λ 는 CV를 통해 얻은 최적값)

λ의 값 ↑	$\lambda = 0$	λ의 값 ↓
λ 영향력 ↑	Penalty term이	λ 영향력↓
$\|\beta\|_1$ ↓	사라지므로	$\|\beta\|_1$ ↑
회귀 계수 ↓	기본 OLS 추정량 도출 !	회귀 계수↑

목적함수에 대한 이해 ②

라그랑지안 승수법을 이용해 나타낸 함수식

$$\text{argmin } \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_j\right)^2 + \lambda\|\beta\|_1$$

λ와 $\|\beta\|_1$의 trade−off 관계

λ의 값 ↑

λ 영향력 ↑

$\|\beta\|_1$ ↓

회귀 계수 ↓

$\lambda = 0$

Penalty term이

사라지므로

기본 OLS 추정량 도출 !

λ의 값 ↓

λ 영향력↓

$\|\beta\|_1$ ↑

회귀 계수↑

목적함수에 대한 이해 ③

Lasso의 목적 함수는 미분 불가능 ▶ closed form이 없음

gradient descent, newton method와 같은
수치적인 방법으로 모수를 추정해야 함

특징

Scaling

개별 변수에 대한 scaling

변수의 단위에 의한 영향력 제거

주로 standard scaling 사용

변수 선택

Ridge달리

λ값이 0이 되는 회귀 계수가 존재하므로

변수 선택이 가능

변수 해석 가능성 증가

예측 성능

상관 관계 높은 변수들이 모델에 존재

예측에 유의미한 변수들을 0으로 만들 수 있음

∴ Ridge에 비해 상대적으로 예측 성능이 떨어짐

특징

Scaling

개별 변수에 대한 scaling

변수의 단위에 의한 영향력 제거

주로 standard scaling 사용

변수 선택

Ridge달리

λ값이 0이 되는 회귀 계수가 존재하므로

변수 선택이 가능

변수 해석 가능성 증가

But,

변수간 상관관계가 높으면 변수 선택 성능이 떨어짐

▼

0이 되는 계수의 존재로 인해 sparsity(희박성)을 지님

예측 성능

상관 관계 높은 변수들이 모델에 존재

예측에 유의미한 변수들을 0으로 만들 수 있음

∴ Ridge에 비해 상대적으로 예측 성능이 떨어짐

특징

Scaling

개별 변수에 대한 scaling

변수의 단위에 의한 영향력 제거

주로 standard scaling 사용

변수 선택

Ridge달리

λ값이 0이 되는 회귀 계수가 존재하므로

변수 선택이 가능

변수 해석 가능성 증가

예측 성능

상관 관계 높은 변수들이 모델에 존재

▼

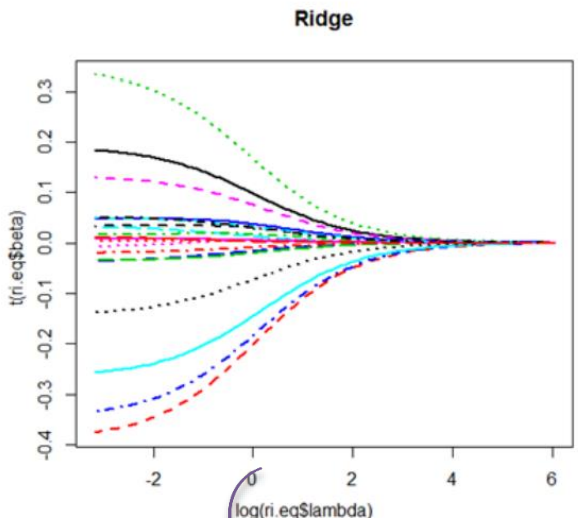
예측에 유의미한 변수들을 0으로 만들 수 있음

∴ Ridge에 비해 상대적으로 예측 성능이 떨어짐

Ridge vs Lasso

Penalty term의 차이

개별 회귀 계수가 변하는 양상에 주목 !



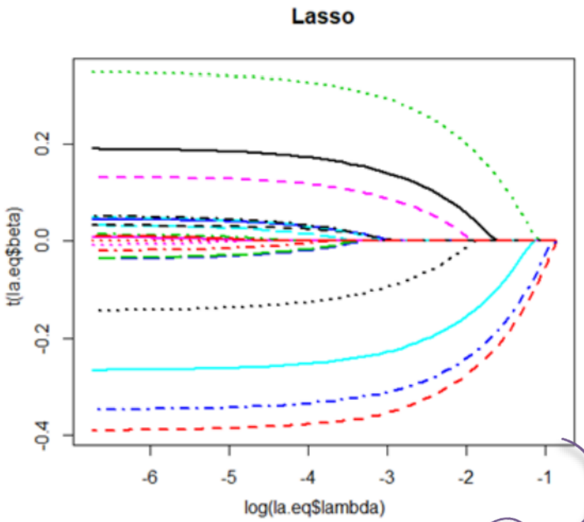
계수가 완만하게 줄어들며

0에 근사할 뿐 0의 값을 가지지 않음

Ridge vs Lasso

Penalty term의 차이

개별 회귀 계수가 변하는 양상에 주목 !



계수가 완만하게 줄어들다 특정 시점 이후 급격히 줄어들며

회귀계수를 정확히 0으로 만드는 λ값 존재

Elastic-Net

Elastic-Net

변수 간 상관관계가 존재할 때 Lasso의 성능이 떨어지는 한계를 보완하기 위한 방법

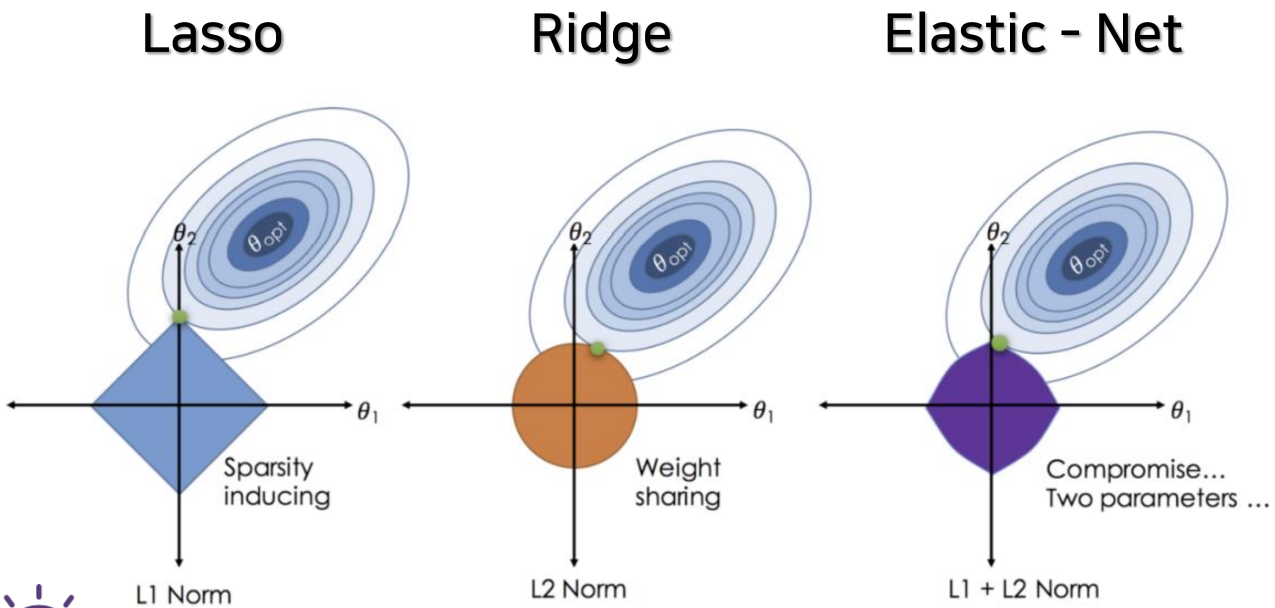
상관성이 있는 변수들을 모두 선택하거나 모두 제거하여 성능 보완

목적함수

$$\operatorname{argmin} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j \right)^2 + \underbrace{\lambda_2 \|\beta\|_2^2}_{\text{Ridge}} + \underbrace{\lambda_1 \|\beta\|_1}_{\text{Lasso}}$$

Penalty term을 혼합

Elastic-Net



제약 조건에 따라 추정량이 만들어지는 공간도 변화

Fused Lasso

Fused Lasso

변수들 사이의 물리적인 거리가 존재한다는 정보를 활용하는 모델

목적함수

$$\operatorname{argmin} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j\right)^2$$

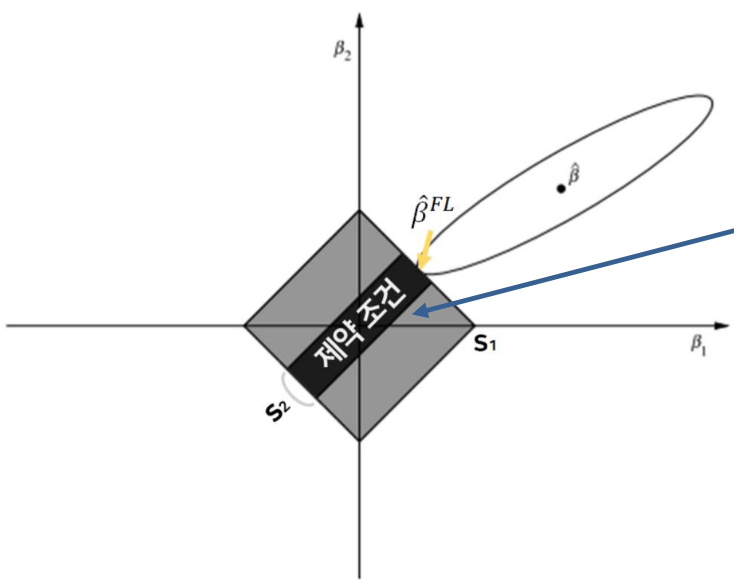
$$\text{subject to } ||\beta||_1 \leq s_1 \text{ and } \sum_{j=2}^{p} |\beta_j - \beta_{j-1}| \leq s_2$$

Lasso
+
인접한 변수들의 회귀 계수를 비슷한 값으로 추정하게 만드는 Penalty term

Fused Lasso

Fused Lasso

변수들 사이의 물리적인 거리가 존재한다는 정보를 활용하는 모델



변수들의 차이가 일정량을 넘어가면
안된다는 제약 조건

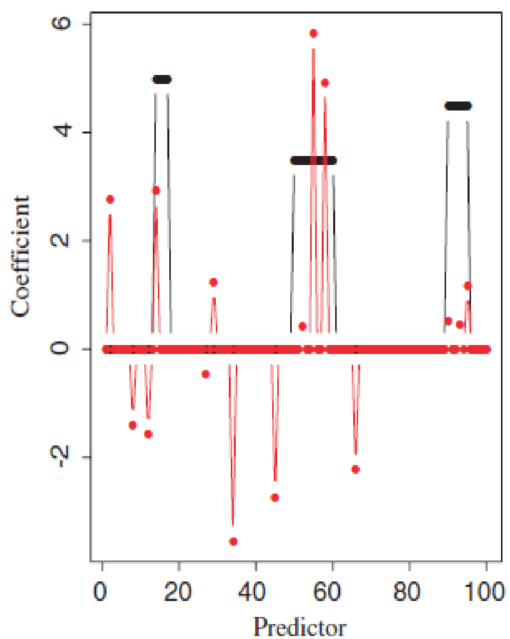
인접한 변수들의 값을 비슷한 값으로 추정

Fused Lasso vs Lasso

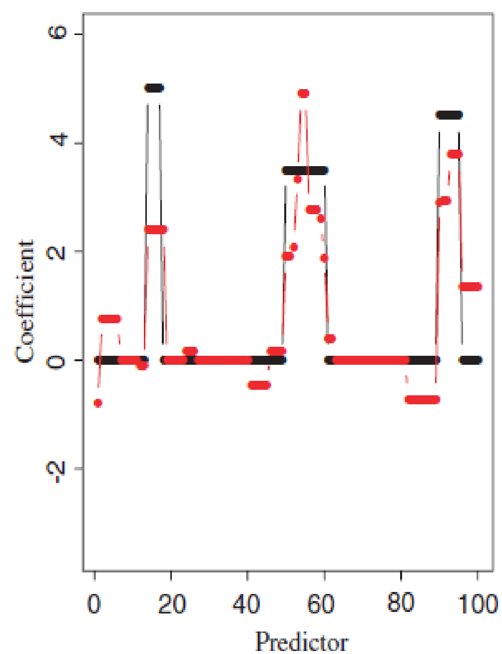
Fused Lasso

변수들 사이의 물리적인 거리가 존재한다는 정보를 활용하는 모델

Lasso



Fused Lasso

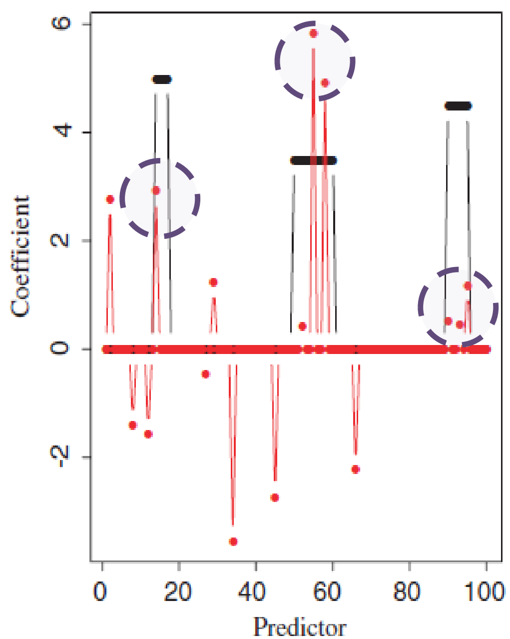


Fused Lasso vs Lasso

Fused Lasso

변수들 사이의 물리적인 거리가 존재한다는 정보를 활용하는 모델

Lasso



Lasso 의 경우

인접한 변수들의 값을

비슷하게 추정하지 못함

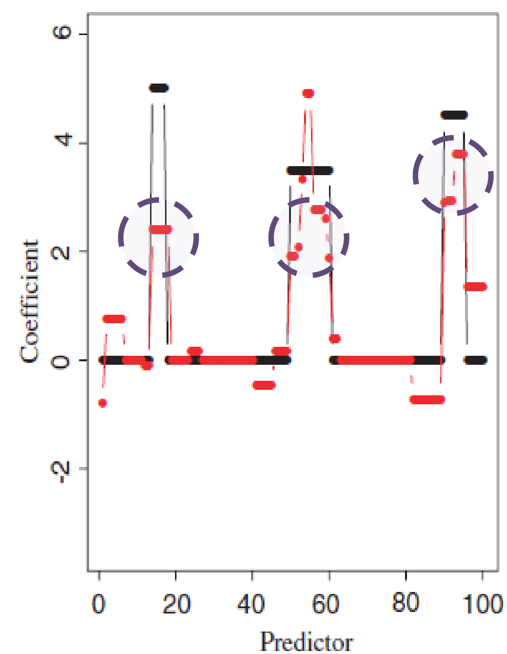
Fused Lasso vs Lasso

Fused Lasso

변수들 사이의 물리적인 거리가 존재한다는 정보를 활용하는 모델

Fused Lasso 의 경우
인접한 변수들의 값을
비슷하게 추정함

Fused Lasso



4

예고


PSAT STUDIOS
회귀
와
육 링즈의 전설
주제분석
10월 대개봉

THANK YOU