회귀분석팀

6팀

조수미
김민지
손재민
박윤아
조웅빈

CONTENTS

다중공선성

다중공선성 Multicollinearity

설명변수 X_j 간에 서로 **선형적인 상관관계**가 존재

즉 설명변수가 서로 간의 선형결합으로 표현 가능

↓

설명변수 간 **독립적**이어야 한다는 가정을 위배

회귀의 기본 가정에 관한 내용은 2주차 클린업 내용 참고!

다중공선성의 문제점

모델의 문제

모델의 검정 결과를 신뢰할 수 없게 됨

전체 회귀식(F-test)은 유의한데
개별 변수 중 유의한 것이 없는
말도 안되는 상황 발생 가능

회귀계수들의 분산이 커짐에 따라 t검정통계량이 작아지기 때문

해석의 문제

개별 회귀계수 β에 관한 해석의 어려움 발생

다중공선성이 발생하면 x_j의 변화가
다른 설명변수를 변화시키므로
나머지 변수가 고정된 상황을 가정하기 힘들어짐

회귀계수 β_j는 설명변수 x_j를 제외한 변수들이 고정되어 있을 때 x_j가 종속변수 Y에 미치는 영향으로 해석

다중공선성 진단 | ③ VIF

VIF 분산팽창인자

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

R_j^2 : $\mathbf{x_j = \gamma_1 x_1 + \cdots + \gamma_{j-1} x_{j-1} + \gamma_{j+1} x_{j+1} + \cdots + \gamma_p x_p}$

다중선형회귀모델을 적합했을 때의 결정계수

일반적으로 VIF_j **값이 10이상인 경우** $(R_j^2 \geq 0.9)$

심각한 다중공선성이 존재한다고 판단 가능

1 다중공선성

다중공선성 해결

Variable Selection

Normalization

Dimension Reduction

Filtering

변수선택법

변수선택법 Variable Selection

수 많은 변수들 중 **적절한 변수 조합**을 찾아내는 방법

서로 상관이 있는 독립변수들을 일부 제거하여 **다중공선성을 해결**

↓

변수선택법을 통해 다중공선성을 완벽하게 제거할 수 없음

그러나 **최종 모델에 대한 확신**을 얻을 수 있음

변수선택지표 | ① Partial F-test

Partial F-test

비교하는 두 모델이 Nested되어 있을 때만 사용 가능

유의하지 않은 변수들을 없애는 방식으로 변수 선택

	Model A	Model B
Nested O	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
Nested X	$y = \beta_0 + \beta_1 x_1 + \beta_4 x_4$	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

Model A $\subset$ Model B 처럼 변수들 집합의 **포함관계 성립**

변수선택지표 | ② 수정결정계수

수정결정계수 adjusted R-squared

설명력을 담당하는 **결정계수**와 **변수 개수 패널티**가

수정결정계수를 구하는 식에 들어감

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

변수선택지표

k는 모델의 모수 개수로, **변수 개수에 따른 패널티**를 부과한 것

AIC Akaike Information Criterion

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

$Likelihood$는 값이 커질수록 모델이 데이터를 잘 설명함

$likelihood$가 커지면 AIC는 작아지는데 AIC가 낮을수록 더 좋은 모형

BIC Bayesian Information Criterion

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

AIC와 다르게 데이터의 개수를 모수의 개수에 곱함으로써 AIC보다 더 큰 패널티를 부과

BIC 역시 낮을수록 더 좋은 모형

변수선택법 | ① Best Subset Selection

Best Subset Selection

가능한 모든 변수들의 조합을 다 고려하는 방법

All Possible Regression

Best Subset Selection Algorithm

1) $\mathrm{M}_1, \cdots, \mathrm{M}_\mathrm{p}$개의 모형 적합

2) (M_1~M_p) p개의 모형 중 AIC 또는 BIC가 가장 작은 모형 선택

3) 만약 AIC/BIC가 가장 작은 모형이 서로 다를 경우 다른 근거에 의해 선택

$M_k(k = 1,2, \ldots, p)$란 변수의 개수를 k개로 적합했을 때 적합한 회귀식 중 MSE가 제일 작은 식

변수선택법 | ② 전진선택법

전진선택법 Forward Selection

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

Forward Selection Algorithm

1) Null Model에서 시작해 X_1부터 X_p까지의 변수들 중 AIC와 BIC를 낮추는 변수를 선택해 추가

2) 만약 X_1이 선택되면 $y = \beta_0 + \beta_1 x_1$의 식에서 X_2부터 X_p까지의 변수들 중 AIC와 BIC를 낮추는 변수 추가

3) 위 과정을 반복하며 AIC와 BIC가 더 이상 낮아지지 않으면 중단

변수선택법 | ③ 후진제거법

후진제거법 Backward Elimination

Forward selection의 반대

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

변수를 하나씩 제거하는 방법

Backward Selection Algorithm

1) Full Model에서 시작해 X_1부터 X_p까지의 변수 중

가장 AIC와 BIC를 낮추는 변수를 선택해 제거

2) 위 과정을 반복하며 AIC와 BIC가 더 이상 낮아지지 않으면 중단

변수선택법 | ④ 단계적 선택법

단계적 선택법 Stepwise Selection

Forward Selection과 Backward Elimination 과정을 섞은 방법

Stepwise Selection Algorithm

1) Forward selection 과정을 통해 가장 유의한 변수들을 모델에 추가

2) 나머지 변수들에 대해 Backward Elimination을 적용

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

정규화

정규화 Regularization

회귀계수가 가질 수 있는 값에 **제약 조건**을 부여함으로써
계수들을 작게 만들거나 0으로 만드는 방법

↓

다중공선성은 OLS 추정량의 분산을 크게 증가시킴
정규화는 OLS 추정량의 **불편성 포기**
But **분산을 줄이는 효과**가 있음

Ridge

Ridge L2 Regularization

SSE를 최소화하면서 회귀계수 β에 제약 조건을 거는 방법

제약 조건식 L2-norm 형태 → L2 Regularization

L2-norm은 선형대수학 2주차 클린업 참고!

목적함수

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

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

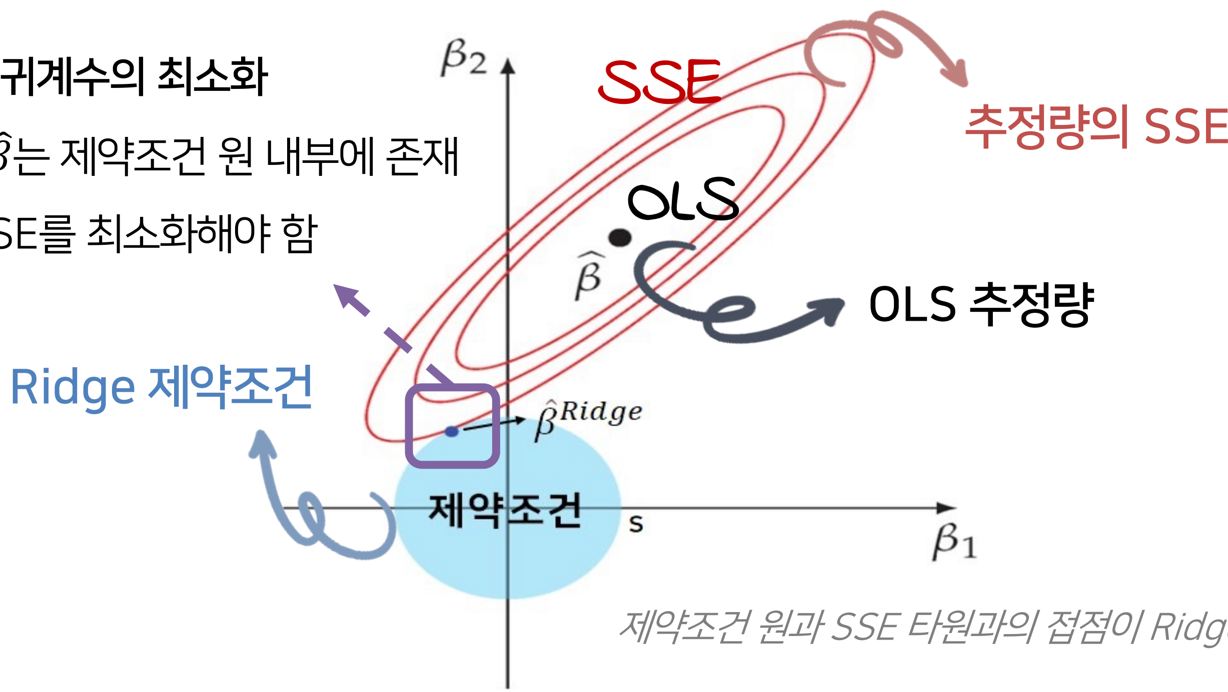
목적함수

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

회귀계수의 최소화

① 회귀계수 $\hat{\beta}$는 제약조건 원 내부에 존재

② SSE를 최소화해야 함



제약조건 원과 SSE 타원과의 접점이 Ridge Estimatior

목적함수 | 라그랑지안 승수법

regularization term을 통해
회귀계수 크기 조정

Lagrangian

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

오차제곱합(SSE) 최소화

음수가 아닌 튜닝 파라미터

최적의 모델을 찾는 과정에서 직접 CV를 통해 조정해주는 모수

제약조건의 크기를 결정 (s와는 반대 관계)

목적함수 | 라그랑지안 승수법

regularization term을 통해 회귀계수 크기 조정

Lagrangian

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

오차제곱합(SSE) 최소화

음수가 아닌 튜닝 파라미터

최적의 ... 찾는 과정에서 직접 CV를 통해 지정해주는 값

제약조건의 크기를 결정 (s와는 반대 관계)

[Λ가 커지는 경우]

λ 의 영향력이 증가하므로

$\sum_{j=1}^{p}\beta_j^2$ 은 작아져야 함

∴ 개별 회귀계수 작아짐

$\Lambda \rightarrow \infty$ 이면, 개별 회귀계수 ≈ 0

[Λ가 작아지는 경우]

λ 의 영향력이 감소하므로

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

∴ 개별 회귀계수 커짐

$\lambda \rightarrow 0$ 이면, OLS 추정량과 동일

Lasso

Lasso L1 Regularization

SSE를 최소화하면서 회귀계수 β에 제약 조건을 거는 방법

제약 조건식 L1-norm 형태 → L1 Regularization

L1-norm은 선형대수학 1주차 클린업 참고!

목적함수

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

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

목적함수

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

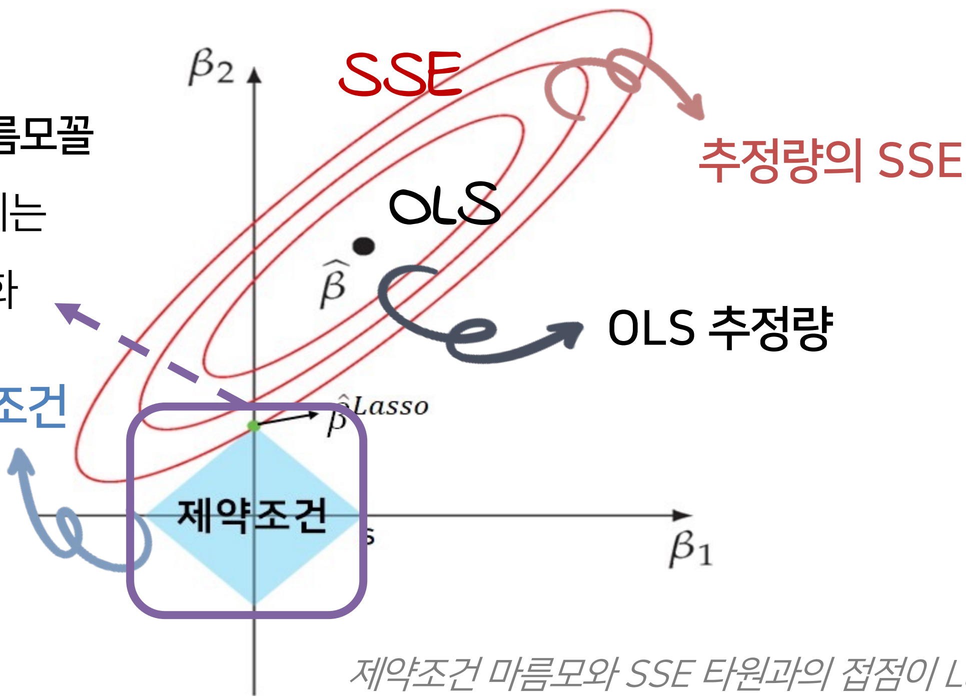
Ridge와 달리

제약 조건의 형태가 마름모꼴

제약조건을 만족시키는

동시에 SSE 최소화

Lasso 제약조건



제약조건 마름모와 SSE 타원과의 접점이 Lasso Estimatior

Elastic-Net

Elastic-Net

상관성이 있는 변수를 모두 제거하거나 선택하여 성능 보완(Grouping Effect)

변수 간 상관관계가 존재할 때

LASSO의 **성능이 떨어지는 한계를 보완**하기 위한 방법

목적함수

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

제약식에 RIDGE의 L2 term과 LASSO의 L1 term이 모두 반영된 모형

Fused Lasso

Fused Lasso

변수들 사이의 물리적 거리가 존재한다는

사전지식을 활용한 모델

목적함수

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

LASSO

New term

인접한 변수들의 회귀 계수를 비슷하게 추정하도록 만드는 역할