

회귀분석팀

6팀

김보근

김민주

서유진

하희나

INDEX

1. INTRO

2. 변수선택법

3. 정규화

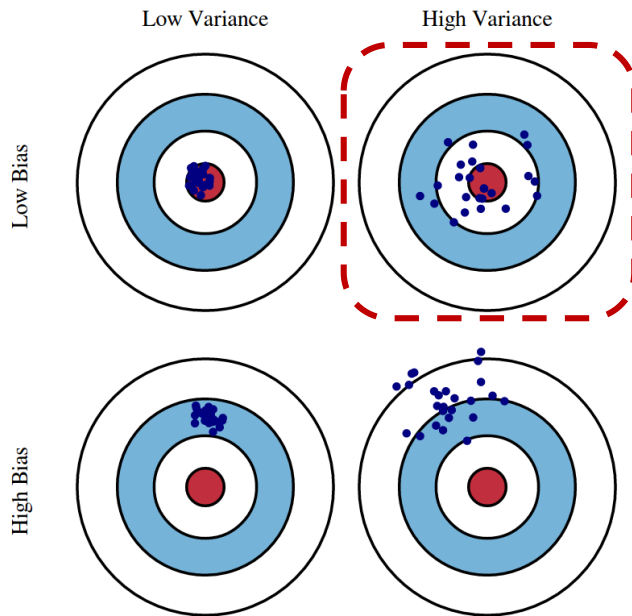
4. 공간회귀분석

1

INTRO

Remind : 다중공선성의 문제

다중공선성은 OLS 추정량의 분산을 크게 증가시킴



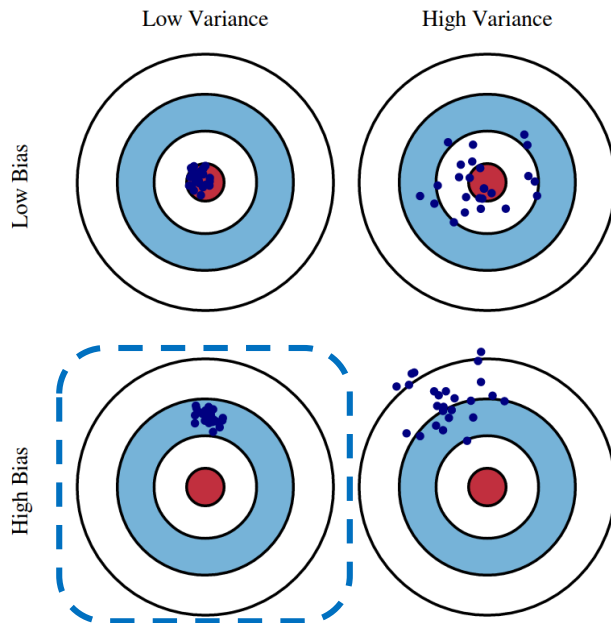
다중공선성이 존재하는 경우
회귀계수 OLS 추정량의 분산 \uparrow



예측 불안정

Remind : 다중공선성의 문제

다중공선성은 OLS 추정량의 분산을 크게 증가시킴



Bias를 조금 포기하더라도
Variance의 감소폭을 더 줄일 수 있다면,
Expected MSE 감소

Remind : 다중공선성의 문제



다중공선성은 OLS 추정량의 분산을 크게 증가시킴

변수 중 일부만을 사용하거나, β 계수를 축소함으로써

bias를 조금 증가시키더라도 분산을 줄이는 것이

다음에 등장할 방법들의 기본적 아이디어



Bias를 조금 포기하더라도

Variance의 감소폭을 더 줄일 수 있다면,

Expected MSE 감소

해결 방법

선대팀 클린업 3주차 참고!

차원축소
(Dimension Reduction)

변수선택법
(Variable Selection)

정규화
(Regularization)

필터링 방법
(Filtering Method)

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

해결 방법

선대팀 클린업 3주차 참고!

차원축소
(Dimension Reduction)

변수선택법
(Variable Selection)

정규화
(Regularization)

필터링 방법
(Filtering Method)

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

2

변수선택법

변수선택법이란?

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!

변수선택법이란?



변수선택법은 **다중공선성이 발견되지 않더라도** 사용 가능!

분석을 위해 고려할 많은 변수들 중 적절한 변수의 조합을 찾아내는 방법

- ▶ 변수 선택을 통해 모델에 대한 **해석력 증가**
우리에게 주어진 후보 변수들(Candidate Regressor) 중에서,
- ▶ 최종 모델에 대한 **확신 증가**
일부분만 중요하거나 예측에 유용할 수 있음

- ▶ 최대한 적은 변수를 사용해 **모형의 분산 감소**



높은 상관관계를 가지는 변수를 제거하여 **다중공선성 해결!**
변수선택법으로 다중공선성을 완벽히 제거하지는 못할 수 있다는 점도 명심!

변수 선택 지표 | ① Partial F-test

일부 회귀계수 group에 대한 유의성 검정

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

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



유의하지 않은 변수들을 없애는 방식으로 변수 선택!

변수 선택 지표 | ① Partial F-test

일부 회귀계수 group에 대한 유의성 검정

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

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

→ Partial F-test **사용 불가**



일반적인 상황에서도 (내포관계와 무관하게)
모델 간의 비교를 가능하게 해주는 지표가 필요

변수 선택 지표 | ② 수정결정계수 (R_{adj}^2)

설명력을 담당하는 결정계수, 변수 개수에 대한 페널티 복합적으로 고려 가능



R_{adj}^2 계산식

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

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

변수 선택 지표 | ③ AIC (*Akaike Information Criterion*)

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

Likelihood 가 커지면 AIC는 작아짐

▶ AIC가 낮을수록 더 좋은 모형으로 해석!

p : 모델의 모수 개수

▶ 변수의 개수에 따른 페널티 부과!

변수 선택 지표 | ③ AIC (*Akaike Information Criterion*)

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

Likelihood 가 커지면 AIC는 작아짐

▶ AIC가 낮을수록 더 좋은 모형으로 해석!

p : 모델의 모수 개수

▶ 변수의 개수에 따른 페널티 부과!

변수 선택 지표 | ③ AIC (*Akaike Information Criterion*)

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

Likelihood 가 커지면 AIC는 작아짐

▶ AIC가 낮을수록 더 좋은 모형으로 해석!



정규분포 따르는 경우

p : 모델의 모수 개수

▶ 변수의 개수에 따른 페널티 부과!

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

변수 선택 지표 | ④ BIC (*Bayesian Information Criterion*)

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

p : 모델의 모수 개수 / n : 데이터의 개수

AIC와 마찬가지로

▶ BIC가 낮을수록 더 좋은 모형으로 해석!

변수의 개수에 데이터의 개수를 곱해

AIC보다 더 큰 페널티를 부과

변수 선택 지표 | ④ BIC (*Bayesian Information Criterion*)

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

p : 모델의 모수 개수 / n : 데이터의 개수

AIC와 마찬가지로

▶ BIC가 낮을수록 더 좋은 모형으로 해석!

변수의 개수에 데이터의 개수를 곱해
AIC보다 더 큰 페널티를 부과

변수 선택 지표 | ④ BIC (*Bayesian Information Criterion*)

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

p : 모델의 모수 개수 / n : 데이터의 개수

AIC와 마찬가지로

▶ BIC가 낮을수록 더 좋은 모형의 해를



정규분포 따르는 경우

변수의 개수에 데이터의 개수를 곱해

AIC보다 더 큰 페널티를 부과

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

변수 선택 방법

Best Subset Selection

가능한 모든 변수들의 조합을 다 고려하는 방법

→ 변수의 개수가 p 개라면, 2^p 개의 모형을 모두 적합하고 비교

비교는 앞서 다뤘던 평가 지표를 통해!

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

▶ $p > 40$ 인 경우 계산 불가능
▶ 적당한 p 에서도 관측치가 많을 경우
계산 비용이 많이 소모

변수 선택 방법

전진선택법

Null Model ($y = \beta_0$) 에서 시작해, **변수**를 하나씩 **추가**하는 방법

후진제거법

Full Model ($y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p$) 에서 시작해,
변수를 하나씩 **제거**하는 방법

단계적 선택법

전진선택법과 후진선택법 과정을 **섞은 방법**



변수 선택 방법

전진선택법

경험적(Heuristic) 방법의 한계

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

▶ 위 방법 모두 Best Subset Selection에 비해 빠르지만,

후진제거법 그럼에도 계산 비용이 굉장히 많이 소모

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

▶ 모든 조합을 고려하지 않아 최고의 모델이라고 확신할 수 없음

▶ 전진선택법과 후진선택법의 결과를 고려했을 때, 둘의 결과가 상이할 수 있음

전진선택법과 후진선택법 과정을 섞은 방법



변수 선택 방법

전진선택법

경험적(Heuristic) 방법의 한계

기계적으로 변수를 **추가** 혹은 **제거**하는 행위는 **매우 위험**



정규화 방법!

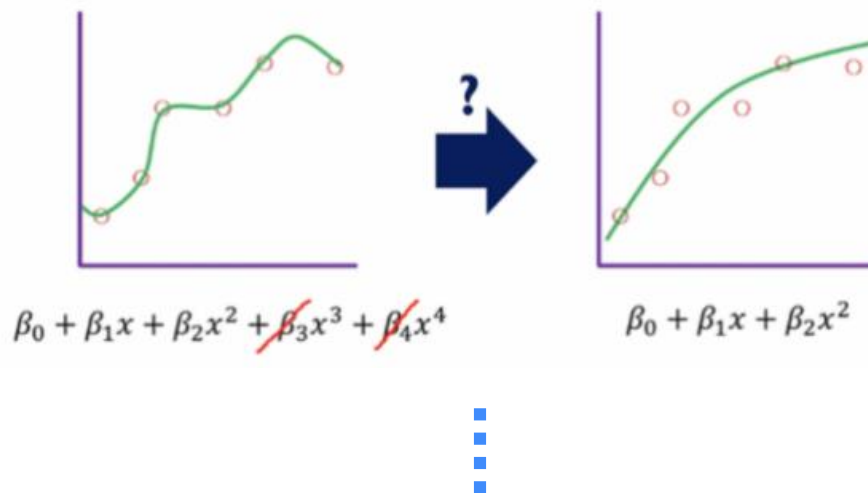
3

정규화(Regularization)

정규화란?

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

EXAMPLE



과적합 방지를 위해 의미 없는 계수가 주는 영향을 줄여보자!

정규화란?

목적함수를 이렇게 짤다면?

$$\min_{\beta} \sum_{i=1} (y_i - \hat{y}_i)^2 + 10000\beta_3^2 + 10000\beta_4^2$$



β_3, β_4 가 조금만 증가해도 값이 큰 폭으로 증가

의미가 없는 변수(β_3, β_4)에
페널티를 부여



목적함수를 만족하도록
 $\beta_3 \approx 0, \beta_4 \approx 0$ 이 되도록 계수 형성

정규화란?

목적함수를 이렇게 짤다면?

$$\min_{\beta} \sum_{i=1} (y_i - \hat{y}_i)^2 + 10000\beta_3^2 + 10000\beta_4^2$$



일반화된 수식으로 확장해보자!

의미가 없는 변수(β_3, β_4)에
페널티를 부여



목적함수를 만족하도록
 $\beta_3 \approx 0, \beta_4 \approx 0$ 이 되도록 계수 형성

정규화 | 목적함수에 대한 이해

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

Training Accuracy에
해당하는 LSE

Generalization Accuracy
(정규화의 증거)

정규화 | 목적함수에 대한 이해

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

Training Accuracy에
해당하는 LSE

Generalization Accuracy
(정규화의 증거)

⋮

의미가 없는 계수의
영향을 줄여 과적합을 방지!


정규화 | 목적함수에 대한 이해

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



오차제곱합(SSE) 최소화 & Regularization term을 통해
개별 회귀계수의 크기가 너무 많이 커지는 것을 조정

정규화 | 목적함수에 대한 이해

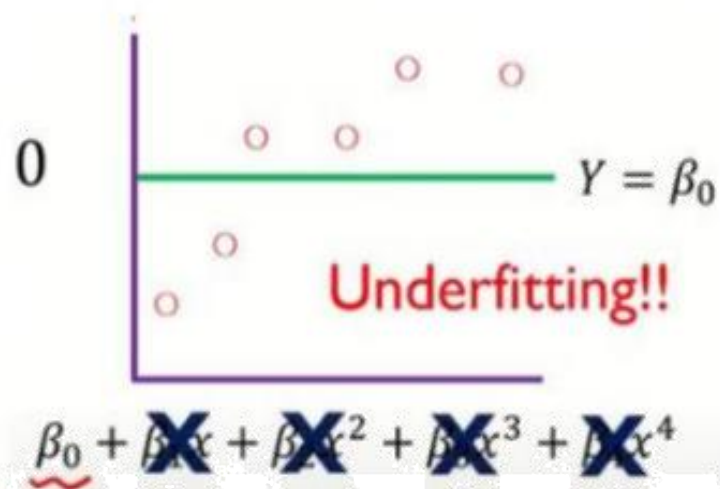
$$L(\beta) = \min_{\beta} \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p \beta_j^2$$


λ 는 우리가 조절할 수 있는 하이퍼파라미터로

LSE와 Generalization Accuracy 사이의 Trade-off를 조절하는 역할

정규화 | 목적함수에 대한 이해

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



▶ λ 가 매우 크다면,

$$\beta_1 \approx 0, \beta_2 \approx 0, \beta_3 \approx 0, \beta_4 \approx 0$$

→ $y = \beta_0$ (직선)

정규화 | 목적함수에 대한 이해

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

▶ λ 가 매우 작다면,
 β 에 대한 제약이 거의 없는 것과 동일



$$\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4$$

3

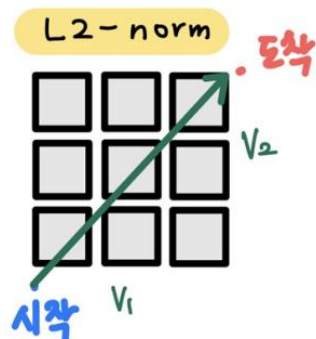
정규화

정규화 | ① Ridge (*L2 Regularization*)

SSE를 최소화하면서 회귀계수 β 에 L2-norm 형태의 제약을 거는 방법

⋮

L2-norm



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

원점에서 벡터까지 연결된 직선 거리

정규화 | ① Ridge (*L2 Regularization*)

목적함수

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

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



위 식을 최소화하여 회귀계수의 Ridge estimator를 얻을 수 있음

정규화 | ① Ridge (*L2 Regularization*)

목적함수

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

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



✓ β 에 대한 이차식 형태이므로 미분을 통해 추정량 계산 가능

✓ 설명 변수들은 표준화된 상태여야 함



정규화 | ① Ridge (L2 Regularization)

목적함수에서 s 와 λ 의 역할

목적함수

s 와 λ 는 정규화를 위한 제약조건이라는 점에서는 같지만

그 영향은 **반대 방향으로 작용** $\text{subject to } \sum_{j=1}^p \beta_j^2 \leq s$

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

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

s 가 작음 = λ 가 큼 \rightarrow 제약을 많이 가함

s 가 큼 = λ 가 작음 \rightarrow 제약을 적게 가함

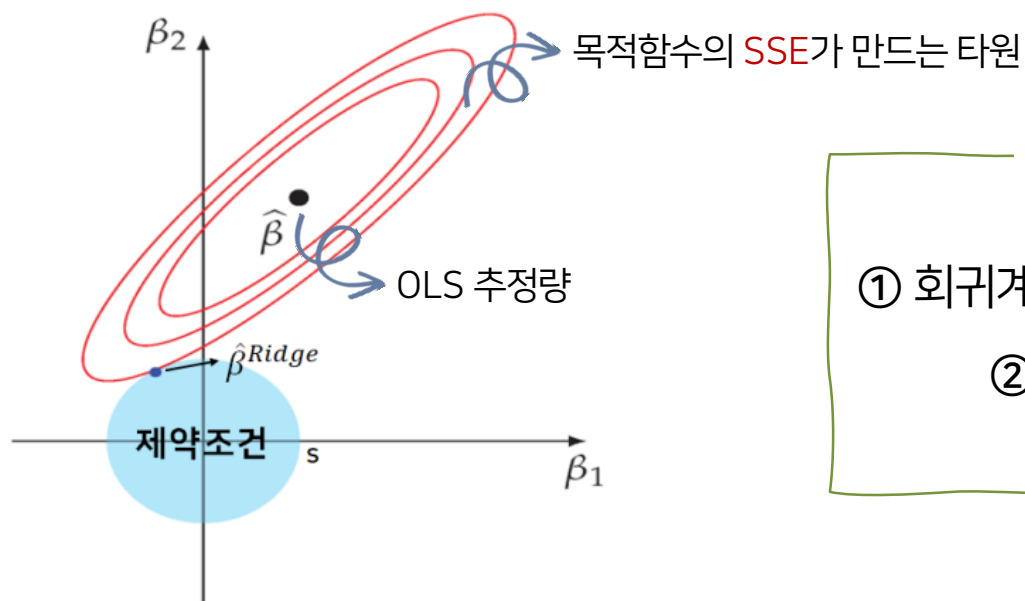
✓ β 에 대한 이차식 형태이므로 미분을 통해 추정량 계산 가능

✓ 설명 변수들은 표준화된 상태여야 함

Ridge | 목적함수에 대한 이해

목적함수

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



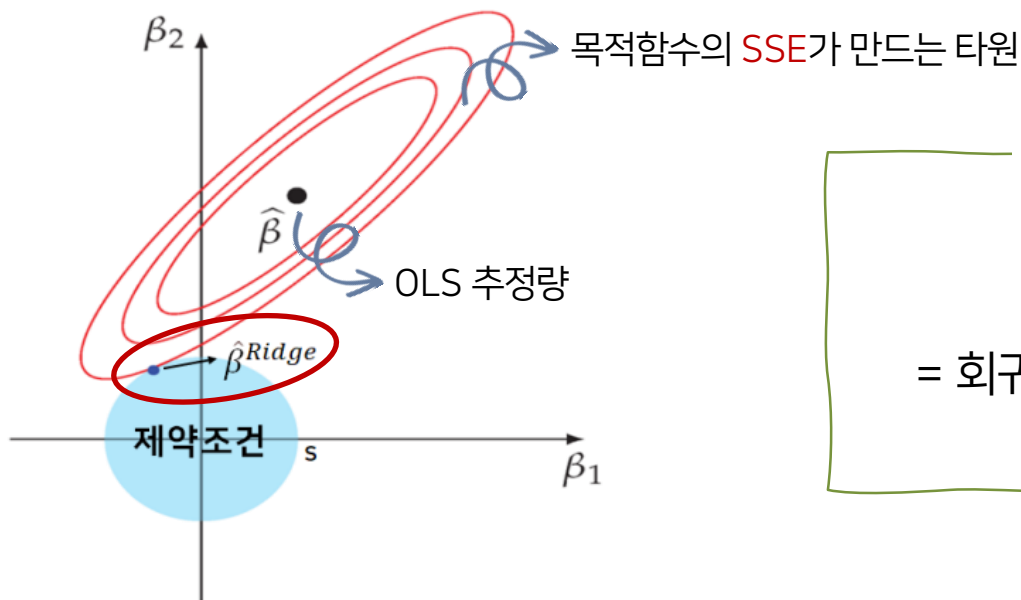
회귀계수의 최소화

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

Ridge | 목적함수에 대한 이해

목적함수

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

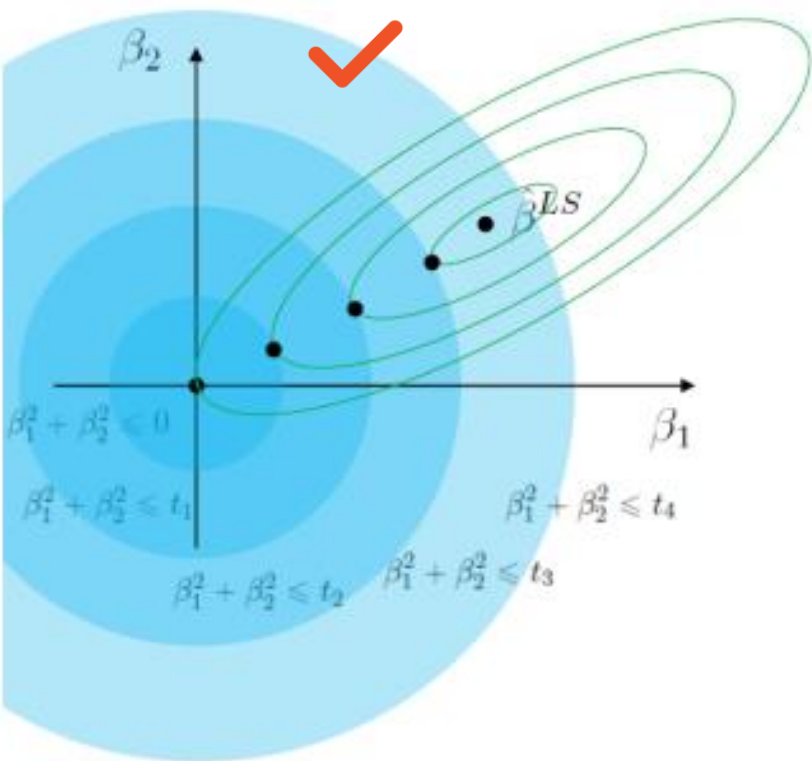


회귀계수의 최소화

타원과 원의 접점

= 회귀계수의 Ridge estimator

Ridge | 목적함수에 대한 이해



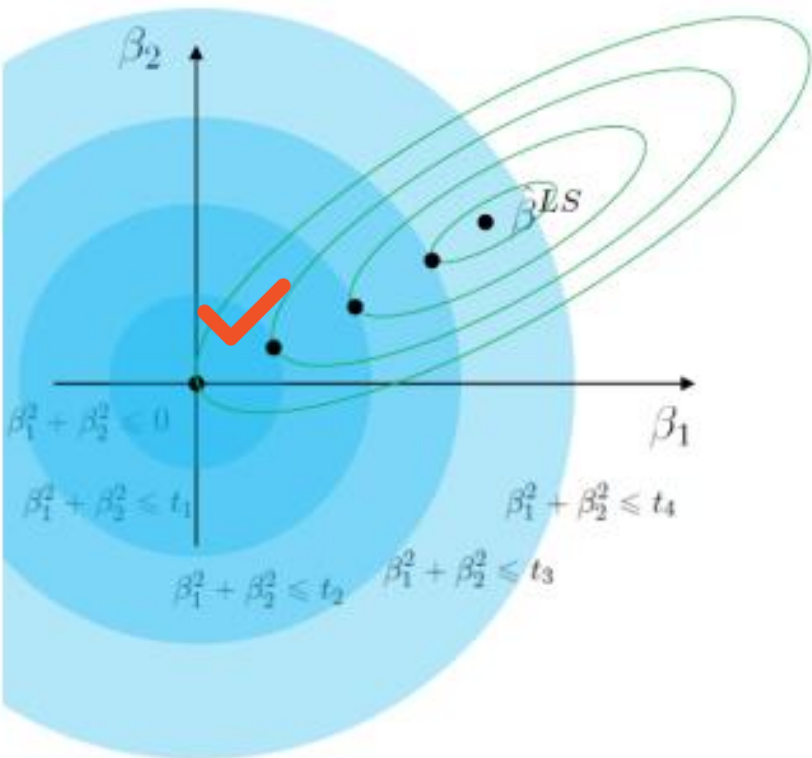
제약조건이 완화된 경우

- ✓ s 가 커질수록 원의 넓이 증가
- ✓ 원이 타원을 밀어내며 추정량이 0에서 멀어짐
- ✓ 회귀계수를 작게 만들 수 없음

제약조건이 강화될 경우

- ✓ s 가 작아지면서 원의 넓이 감소
- ✓ 추정량이 0으로 수렴함(0은 될 수 없음)
- ✓ 회귀계수를 작게 만들 수 있음

Ridge | 목적함수에 대한 이해



제약조건이 완화될 경우

- ✓ s 가 커질수록 원의 넓이 증가
- ✓ 원이 타원을 밀어내며 추정량이 0에서 멀어짐
- ✓ 회귀계수를 작게 만들 수 없음

제약조건이 강화될 경우

- ✓ s 가 작아지면서 원의 넓이 감소
- ✓ 추정량이 0으로 수렴함(0은 될 수 없음)
- ✓ 회귀계수를 작게 만들 수 있음

Ridge | 목적함수에 대한 이해

λ 의 값에 따른 회귀계수의 변화

λ 가 커지는 경우 [λ 의 영향력이 증가]

전체 식을 최소화하기 위해

$\sum_{j=1}^p \beta_j^2$ 은 작아져야 함

▶ 개별 회귀 계수들은 감소

λ 가 작아지는 경우 [λ 의 영향력이 감소]

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

▶ 개별 회귀 계수들은 증가

- ✓ s 가 작아지면서 원의 넓이 감소
- ✓ 추정량이 0으로 수렴함(0은 될 수 없음)
- ✓ 회귀계수를 작게 만들 수 있음

Ridge | 목적함수에 대한 이해

λ 의 값에 따른 회귀계수의 변화

$\lambda \rightarrow \infty$

개별 회귀계수의 영향력은

무시될 만큼 작아짐

▶ 회귀 계수 ≈ 0

$\lambda = 0$

Regularization term이 없어짐

▶ 기존 OLS 추정량과 동일

제약조건이 강화될 경우

✓ s 가 작아지면서 원의 넓이 감소

✓ 추정량이 0으로 수렴함(0은 될

✓ 회귀계수를 작게 만들 수 있다



Ridge | 특징

① Scaling

회귀계수는 변수 단위에 큰 영향을 받음

→ Scaling을 통해 단위의 영향을 제거, 순수 영향력만을 사용

주로 Standard Scaling 사용

② 계산 비용 절약

Regularization Term이 L2 norm 형태이므로 미분 가능

→ λ 를 바꾸며 미분과 함께 행렬 연산 가능

Ridge | 특징

① Scaling

회귀계수는 변수 단위에 큰 영향을 받음

→ Scaling을 통해 단위의 영향을 제거, 순수 영향력만을 사용

주로 Standard Scaling 사용

② 계산 비용 절약

Regularization Term이 L2 norm 형태이므로 미분 가능

→ λ 를 바꾸며 미분과 함께 행렬 연산 가능

Ridge | 특징

③ 예측 성능

상관관계가 높은 변수들이 모델에 존재할 경우 좋은 예측 성능을 보임

④ 변수 선택 불가

다중공선성의 영향력을 줄일 뿐 원인이 되는 변수 제거 불가

→ 해석력 증가는 기대하기 어려움



λ 값이 커지면서 개별 회귀계수가 0에 가까워지기는 하지만, 0이 되지는 않기 때문

Ridge | 특징

③ 예측 성능

상관관계가 높은 변수들이 모델에 존재할 경우 좋은 예측 성능을 보임

④ 변수 선택 불가

다중공선성의 영향력을 줄일 뿐 원인이 되는 **변수 제거 불가**

→ 해석력 증가는 기대하기 어려움



λ 값이 커지면서 개별 회귀계수가 0에 가까워지기는 하지만, 0이 되지는 않기 때문



Ridge | 특징

Ridge Regression 행렬로 이해하기

③ 예측 성능

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

상관관계가 높은 변수들이 모델에 존재할 경우 좋은 예측 성능을 보임

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

④ 변수 선택 불가

$$\hat{\beta}^{ridge} = (X^T X + \lambda I_p)^{-1} X^T y \quad \text{vs} \quad \hat{\beta}^{OLS} = (X^T X)^{-1} X^T y$$

다중공선성의 영향력을 줄일 수 있는 원인이 되는 변수 제거 불가

→ 해석력 증가는 기대하기 어려움

 I_p 는 $p \times p$ Identity Matrix이므로대각요소 각각에 λ 만큼 더한 것과 동일

λ 값이 커지면서 개별 회귀계수가 0에 가까워지기는 하지만, 0이 되지 않는기 때문에
이와 같은 closed form이 존재하기에 계산 비용이 줄어듦



Ridge | 특징

Ridge Regression 행렬로 이해하기

③ 예측 성능

Ridge estimator의 추정량은 $X^T X$ 에 λI_p 를 더해준 꼴

$$\hat{\beta}^{ridge} = (X^T X + \lambda I_p)^{-1} X^T y \quad \text{vs} \quad \hat{\beta}^{OLS} = (X^T X)^{-1} X^T y$$

④ 변수 선택 불가

행렬을 full rank로 만들거나 행렬식을 크게 만들 수 있음

$$\rightarrow \det(X^T X) \leq \det(X^T X + \lambda I_p)$$

λ 값이 커지면서 개별 회귀계수가 0에 가까워지기는 하지만, 0이 되지는 않기 때문

다중공선성 해결 가능!





Ridge | 특징

행렬연산을 통한 Closed Form Solution

③ 예측 성능

상관관계가 높은 변수들이 모델에 존재할 경우 좋은 예측 성능을 보임

OLS는 BLUE에 의해 Unbiased
Ridge는 λ 만큼 더했기에 Biased

④ 변수 선택 불가

다중공선성의 영향력을 줄일 뿐 원인이 되는 변수 제거 불가

→ 해석력 증가에 반대하기 어려움

그러나 Variance가 작기 때문에

더 높은 예측 성능을 가짐

λ 값이 커지면서 개별 회귀계수는 0에 가까워지지만, 0이 되지 않는 이유

짱이다..너무 짱인데..
진짜 짱이다..와..
완전 캡짱이다..



3

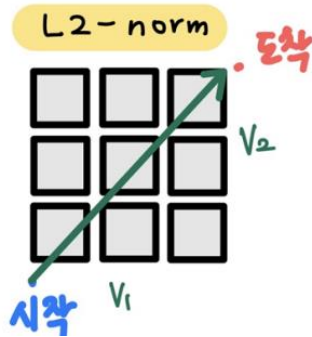
정규화

정규화 | ② Lasso (*L1 Regularization*)

SSE를 최소화하면서 회귀계수 β 에 L1-norm 형태의 제약을 거는 방법

⋮

L1-norm



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

원점에서 벡터까지의 각 좌표의 합

정규화 | ② Lasso (*L1 Regularization*)

목적함수

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

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



위 식을 최소화하여 회귀계수의 Lasso estimator를 얻을 수 있음

정규화 | ② Lasso (*L1 Regularization*)

목적함수

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

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

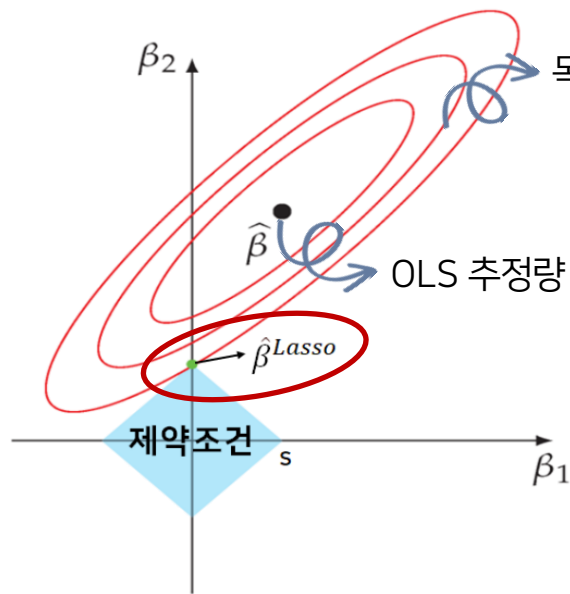


- ✓ 미분이 불가능하므로 수치적인 방법을 이용해 최적화 문제를 해결해야 함
- ✓ 설명 변수들은 표준화된 상태여야 함

Lasso | 목적함수에 대한 이해

목적함수

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



목적함수의 SSE가 만드는 타원

OLS 추정량

 $\hat{\beta}^{Lasso}$

제약조건

s

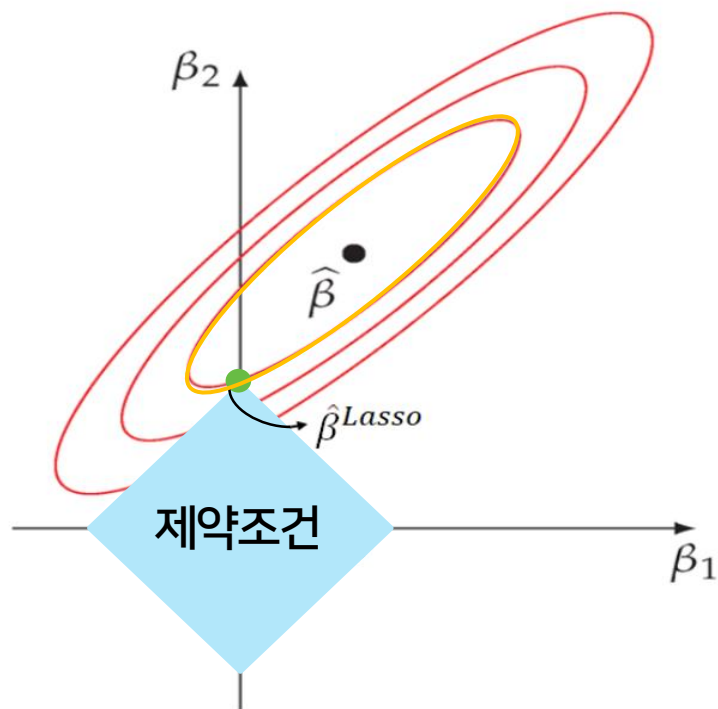
 β_1 β_2

회귀계수의 최소화

타원과 마름모의 접점

= 회귀계수의 Lasso estimator

Lasso | 목적함수에 대한 이해

① s 가 커질 때

마름모의 넓이가 커짐

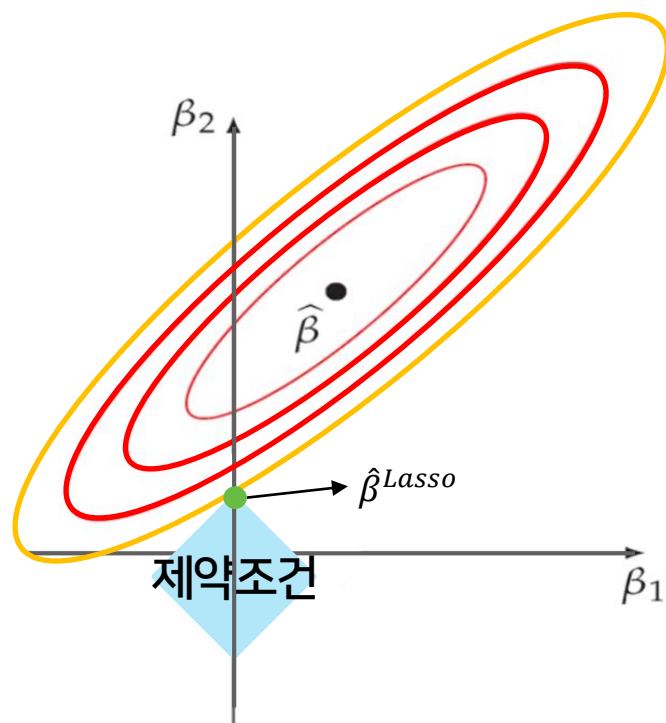


제약 조건이 완화됨



회귀계수 추정량이 0에서 멀어짐

Lasso | 목적함수에 대한 이해

② s 가 작아질 때

마름모의 넓이가 작아짐

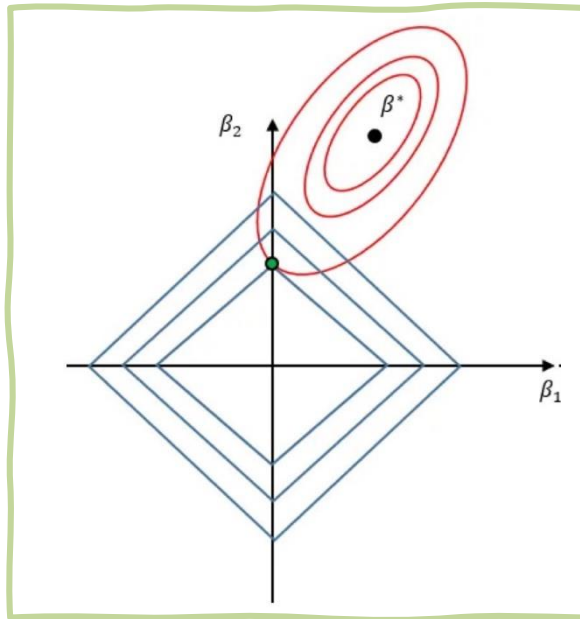


제약 조건이 강화됨

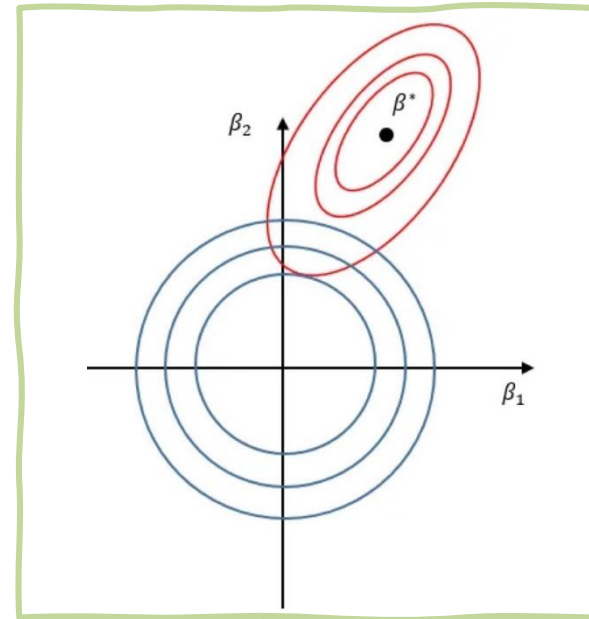


회귀계수 추정량이 0과 가까워짐

Lasso와 Ridge의 제약조건



▲ Lasso

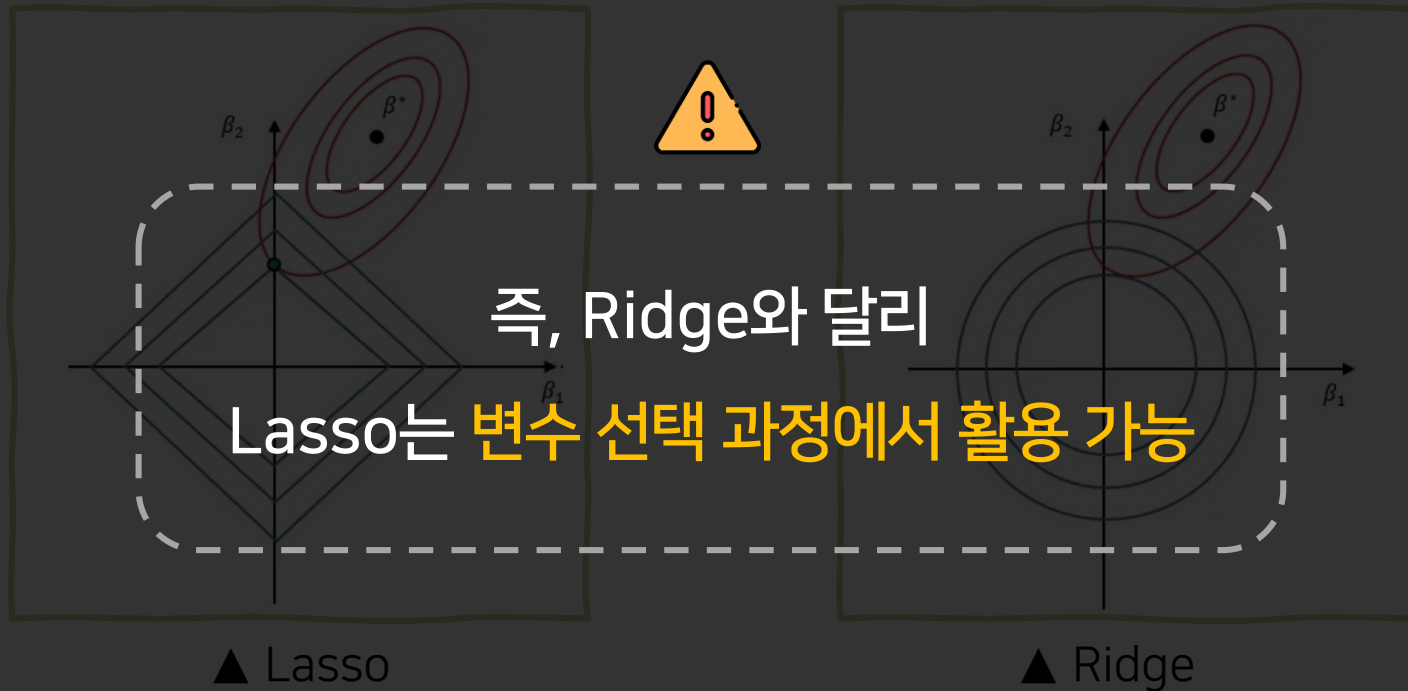


▲ Ridge

Lasso는 제약조건기의 형태로 인해 Ridge와 달리 회귀계수 추정량이 0이 될 수 있음

→ 이를 variable selection 과정에 활용 가능!

Lasso와 Ridge의 제약조건



Lasso는 제약조건기의 형태로 인해 Ridge와 달리 회귀계수 추정량이 0이 될 수 있음

→ 이를 variable selection 과정에 활용 가능!

Lasso | 목적함수에 대한 이해

목적함수

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

라그랑주 승수법으로 변환된 목적함수

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



오차제곱합(SSE) Term



Regularization Term

개별 회귀계수가 너무 커지는 것을 조정

Lasso | 목적함수에 대한 이해

라그랑주 승수법으로 변환된 목적함수

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

 λ 가 커지는 경우

λ 의 영향력이 증가하므로,
전체 식을 최소화하기 위해

$\sum_{j=1}^p |\beta_j|$ 은 작아져야 함

▶ 개별 회귀 계수들은 감소

 λ 가 작아지는 경우

λ 의 영향력이 감소하므로,
상대적으로 $\sum_{j=1}^p |\beta_j|$ 영향력 증가

▶ 개별 회귀 계수들은 증가

Lasso | 목적함수에 대한 이해

라그랑주 승수법으로 변환된 목적함수

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

 λ 가 커지는 경우

$$\lambda \rightarrow \infty$$

개별 회귀계수의 영향력은

무시될 만큼 작아짐

▶ 회귀 계수 ≈ 0 λ 가 작아지는 경우

$$\lambda = 0$$

Regularization term 없어짐

▶ OLS 추정량과 동일

Lasso | 목적함수에 대한 이해



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

Lasso | 특징

① Scaling

회귀계수는 **변수 단위**에 큰 영향을 받음

→ Scaling을 통해 단위의 영향을 제거, **순수 영향력**만을 사용

주로 Standard Scaling 사용

② 변수 선택

0이 되는 회귀 계수가 존재해 변수 선택 가능

→ 변수 선택으로 해석 가능성이 증가하지만

변수 간 상관 관계가 높다면 변수 선택 성능이 떨어짐

Lasso | 특징

① Scaling

회귀계수는 변수 단위에 큰 영향을 받음

→ Scaling을 통해 단위의 영향을 제거, 순수 영향력만을 사용

주로 Standard Scaling 사용

② 변수 선택

0이 되는 회귀 계수가 존재해 변수 선택 가능

→ 변수 선택으로 해석 가능성이 증가하지만

변수 간 상관 관계가 높다면 변수 선택 성능이 떨어짐

Lasso | 특징

③ 예측 성능

변수들 간 상관관계가 큰 경우,
예측에 유의미한 변수들을 0으로 만들 수 있어 예측 성능이 떨어짐

④ Closed form solution

미분 불가능한 점이 있어 closed form solution을 구할 수 없음
→ 수치 최적화 방법 사용

Lasso | 특징

③ 예측 성능

변수들 간 상관관계가 큰 경우,
예측에 유의미한 변수들을 0으로 만들 수 있어 예측 성능이 떨어짐

④ Closed form solution

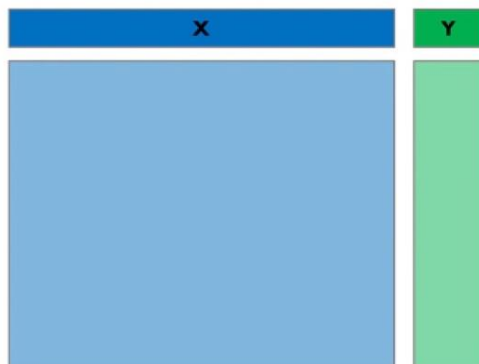
미분 불가능한 점이 있어 closed form solution을 구할 수 없음

→ 수치 최적화 방법 사용

Lasso는 또한 꽤 **강건**한 모델!

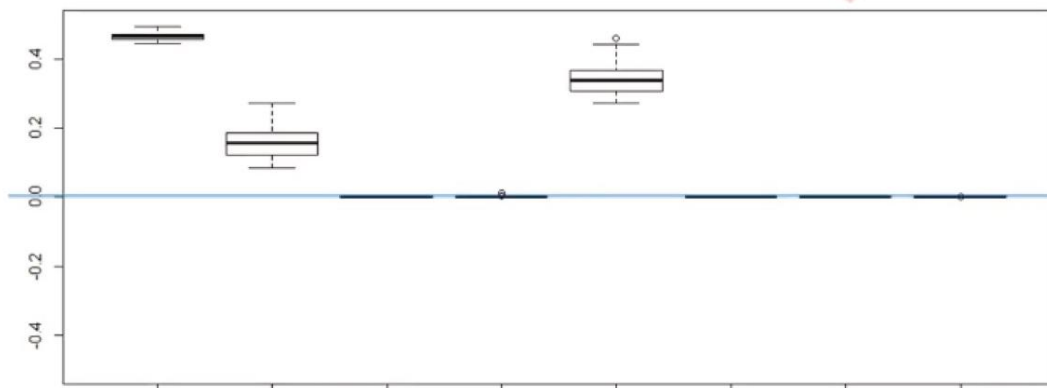
Lasso | 특징

③ 예측



$\hat{\beta}_1, \dots, \hat{\beta}_8$
 $\hat{\beta}_1, \dots, \hat{\beta}_8$
 $\hat{\beta}_1, \dots, \hat{\beta}_8$
 $\hat{\beta}_1, \dots, \hat{\beta}_8$
 $\hat{\beta}_1, \dots, \hat{\beta}_8$

④ Close



Boxplot을 통해 변화 폭이 크지 않고, 0이 된 변수들은 계속 0이 됨을 확인 가능

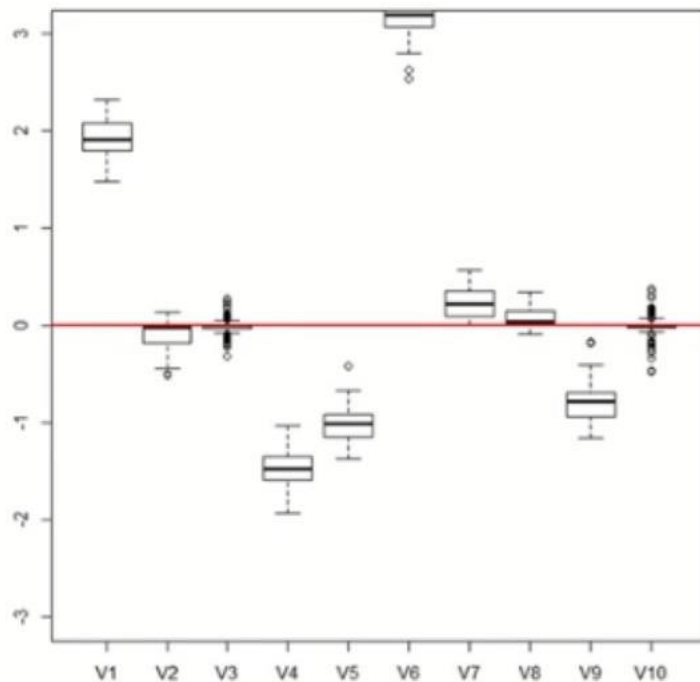
→ 데이터가 바뀌더라도 결과는 강건하다



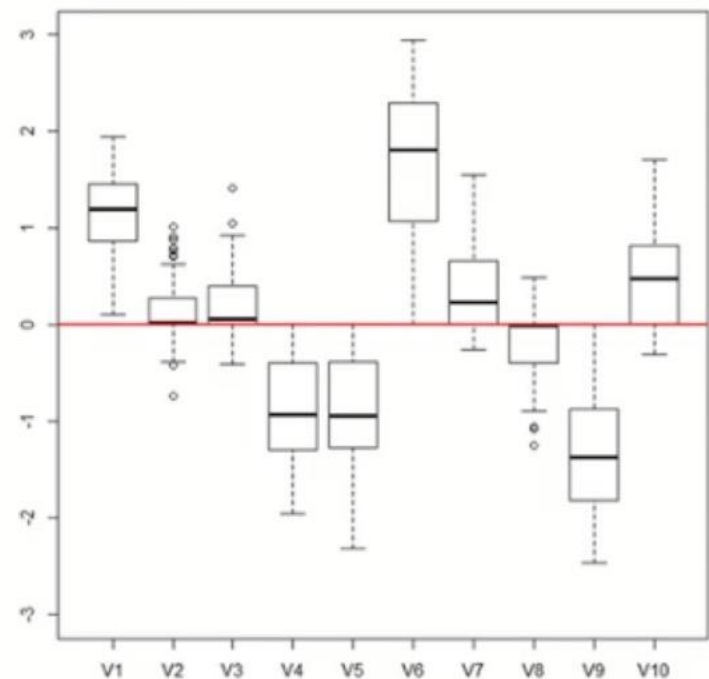
Lasso

그러나 변수 간 상관관계가 높다면 robust하지 않음

변수 간 상관관계가 낮을 경우



변수 간 상관관계가 높을 경우



Ridge vs. Lasso 정리

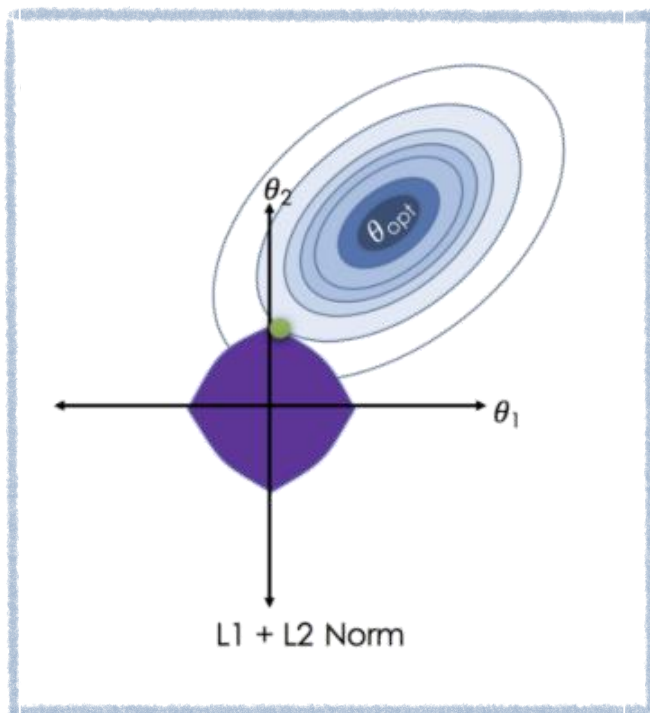


Ridge	Lasso
변수 선택 불가능	변수 선택 가능
Closed Form Solution 0 (미분이 가능함)	Closed Form Solution X (미분이 불가능함)
변수 간 상관관계가 높은 상황에서 좋은 예측 성능	변수 간 상관관계가 높은 상황에서 Ridge에 비해 예측 성능 ↓
제약 범위가 원	제약 범위가 마름모꼴
크기가 큰 변수를 우선적으로 줄임	



Elastic-Net

Ridge와 Lasso의 Regularization term을 혼합한 방법



Elastic Net의 제약조건은
Ridge, Lasso 제약조건의 중간 형태

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 제거하여 성능 보완



Grouping Effect!

Elastic-Net

목적함수

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

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

⋮

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

Ridge L2 Term

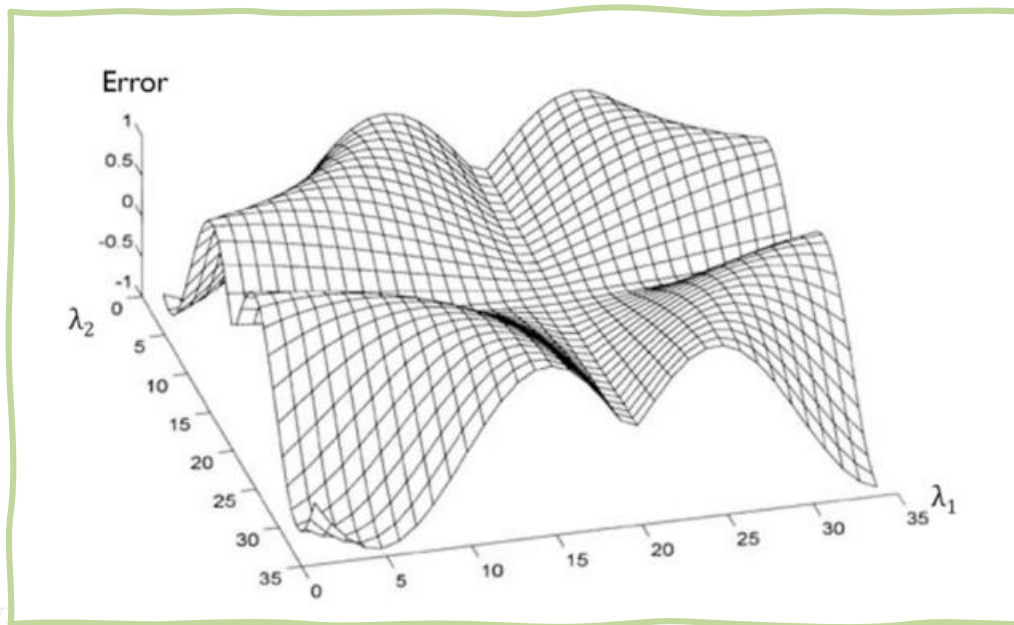
Lasso L1 Term



Elastic-Net

Elastic Net의 Parameter

목적함수



Grid Search 방법을 사용하여 범위 내에서
오차를 최소화하는 λ_1 과 λ_2 의 조합을 찾음

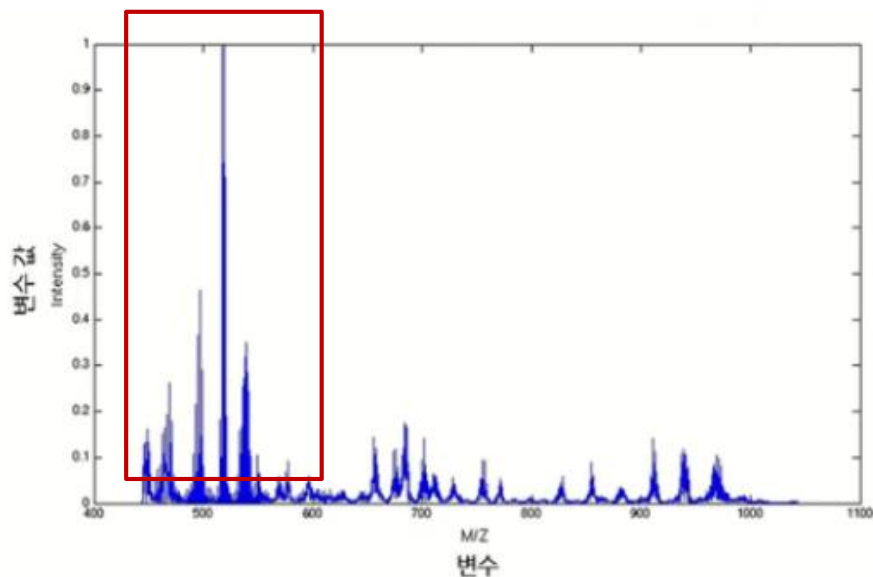
$$\hat{\beta}^{Elastic} = \arg \min_{\beta} \sum_{i=1}^n \ell(\beta_i) + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2$$

Ridge L2 Term

Lasso L1 Term

Fused Lasso

변수들 사이의 인접성에 대한 사전 지식을 이용한 회귀 모델



Signal, Spectra와 같은 데이터는
중요한 변수들이 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

물리적으로 인접한 변수들의 회귀계수를 비슷한 값으로 추정하게 함



양 옆에 위치한 변수들의 회귀계수 값의 차이를 최소화하는 Smoothness 역할



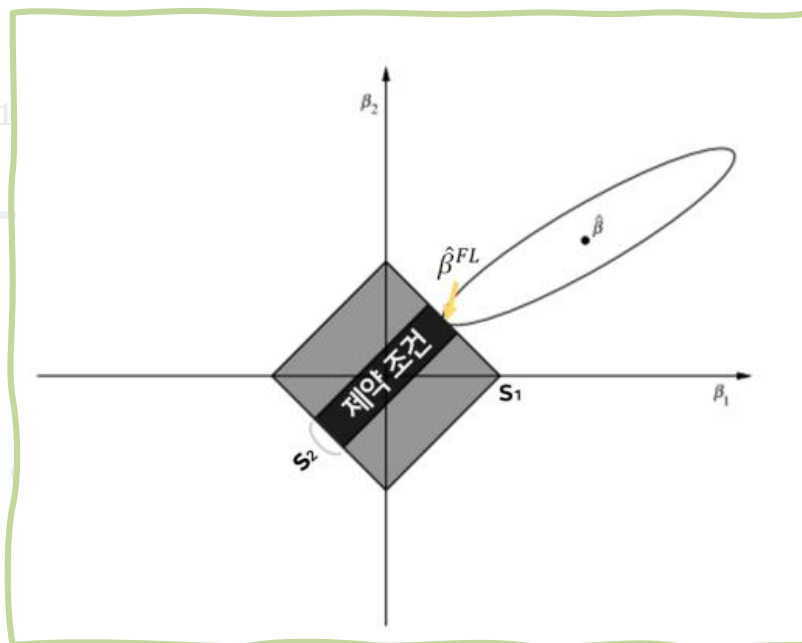
Fused Lasso

Fused Lasso 제약조건

· 목적함수

$$\hat{\beta}^{FL} = \operatorname{argmin}_{\beta} \left(\sum_{i=1}^n \right)$$

$$+ \lambda_2 \sum_{j=1}^p |\beta_j - \beta_{j-1}|$$



물리적으로

추정하게 함

기존 Lasso 제약 공간보다 더 **엄격하게 제약 공간이 형성됨!**

양 옆에 위치한 변수들의 회귀계수 값의 차이를 최소화하는 Smoothness 역할

4

공간회귀분석

2주차 Remind : 오차의 독립성 위배 시 처방 방법

1) 설명변수 추가

자기상관을 유발하는 변수를 설명변수로 모형에 추가

2) 분석 모델 변경

✓ **시간**에 따른 자기상관

→ 자기 상관을 고려하는 AR(p) 같은 **시계열 모델** 사용

✓ **공간**에 따른 자기상관

→ 공간의 인접도를 고려하는 **공간회귀모델** 사용

2주차 Remind : 오차의 독립성 위배 시 처방 방법

1) 설명변수 추가

자기상관을 유발하는 변수를 설명변수로 모형에 추가

2) 분석 모델 변경

✓ **시간**에 따른 자기상관

→ 자기 상관을 고려하는 AR(p) 같은 **시계열 모델** 사용

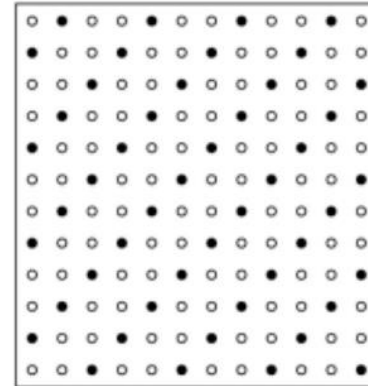
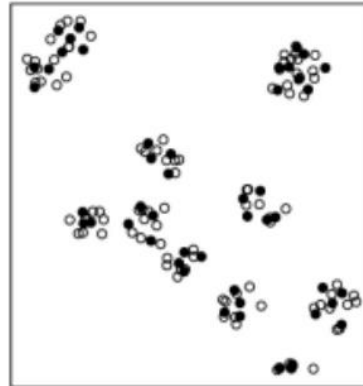
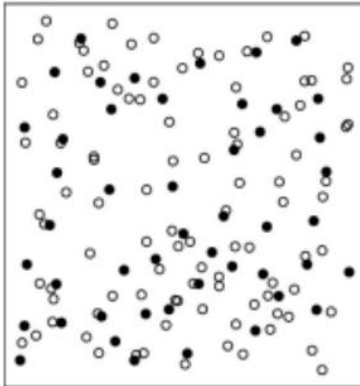
✓ **공간**에 따른 자기상관

→ 공간의 인접도를 고려하는 **공간회귀모델** 사용



공간 데이터

공간 상의 **위치** 또는 **좌표**와 관련된 속성의 집합



공간 패턴 분석

공간 패턴을 형성하는 데 영향을 미친 **공간 과정**을 파악

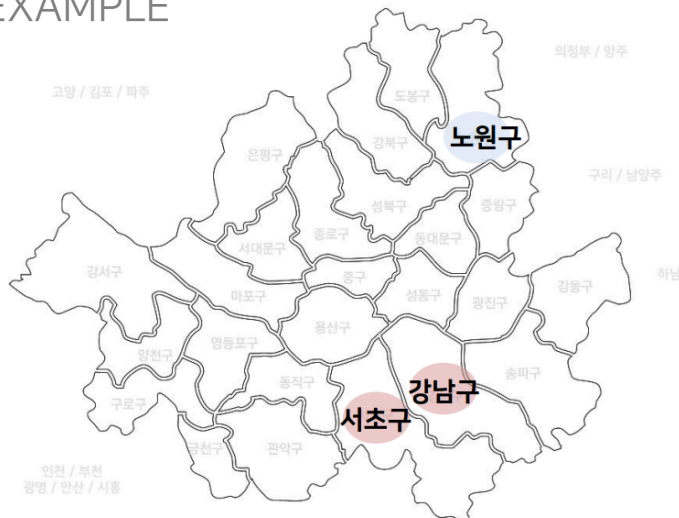
공간자기상관

Tobler 지리학 제 1 법칙

*Everything is related to everything else,
but near things are more related than distant things*

가까이 있을수록 **유사성**을 띠는 공간데이터의 특성을 **공간자기상관**이라 함

EXAMPLE

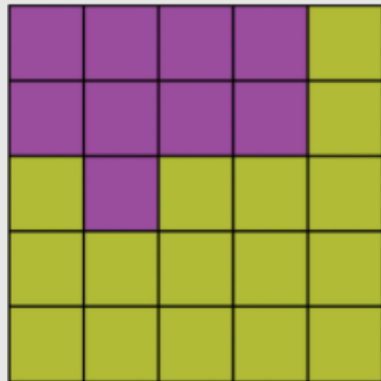


강남구는 노원구보다
지리적으로 **가까운** 서초구와
아파트 가격이 **비슷함**

공간자기상관

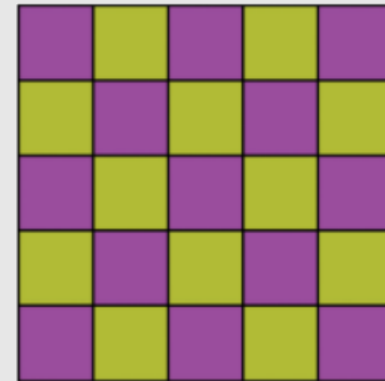
상관 방향에 따른 구분

양의 공간자기상관



근처의 관측치들과 유사한 형태

음의 공간자기상관



근처의 관측치들과 상반된 형태

공간자기상관

공간 크기에 따른 구분

전역적 공간자기상관

전체 구역이 가지는
하나의 공간자기상관 정도

ex) 서울시에서 나타나는
집값의 공간적인 패턴

국지적 공간자기상관

특정 지점이 가지는
개별적인 공간자기상관 정도

ex) 혜화동에서 나타나는
집값의 공간적인 패턴

공간데이터의 특성 ① : 공간적 의존성

가까이 있는 공간 데이터가 유사성을 띠는 것을 의미하며,
인접 지역의 Y가 **근처의 Y에게 영향**을 주는 것을 말함



특정 사건의 강도가 인접 지역의 사건 강도에 영향을 주는가?



EXAMPLE

종로구의 지가 상승이 성북구의 지가를 상승시키는가?

공간데이터의 특성 ② : 공간적 이질성

넓은 지역에서 나타나는 불규칙한 분포를 의미하며,
한 지역 내에 서로 다른 성격의 하위 집단이 존재하는 것을 말함



특정 사건이 전 지역에서 동일한 강도로 나타나는가?



EXAMPLE

지하철 개통이 집값에 미치는 영향력의 크기가 도시, 농촌에서 같은가?
→ 영향을 많이 받는 지역, 영향을 적게 받는 지역 등 여러 유형이 존재 가능

공간자기상관 진단

먼저, 알고자 하는 지역들이 **공간적으로 인접한지**부터 확인해야 함!

공간가중행렬

지역 내의 지점들이 서로 공간적으로 **인접하고 있는지**의 여부를
파악할 수 있도록 **행렬**로 나타낸 것

⋮

$$w_{ij} = \begin{cases} 1 & \text{if } i, j \text{ is neighbor} \\ 0 & \text{otherwise} \end{cases}$$

공간가중행렬의 이웃 결정 기준

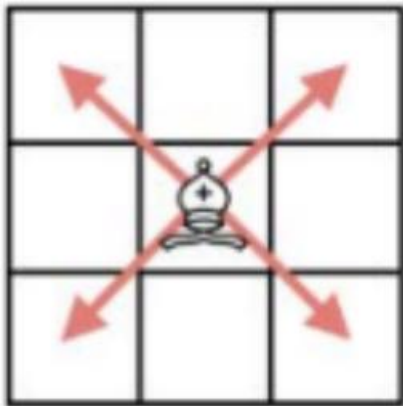
Data에 적합하게 스스로 정의도 가능!

Binary Contiguity Weights	Bishop Contiguity
	Rook Contiguity
	Queen Contiguity
Distance-based Weights	
K-Nearest Neighbors Weights	

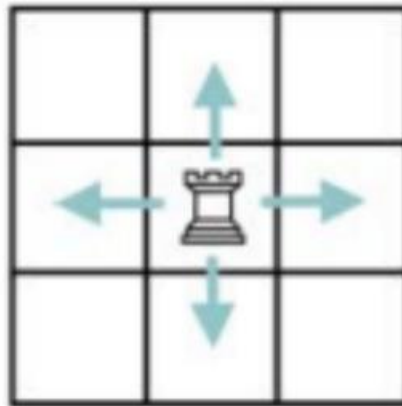
공간가중행렬의 이웃 결정 기준

① Binary Contiguity Weights

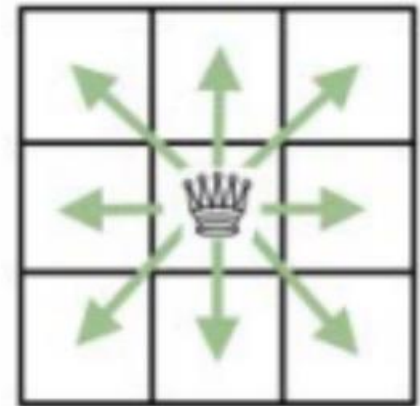
근접하고 있는 경우를 이웃으로 보는 방법



▲ Bishop Contiguity



▲ Rook Contiguity



▲ Queen Contiguity

가장 보편적으로 사용!

공간가중행렬의 이웃 결정 기준

② Distance-based Weights

특정 거리보다 가까우면 이웃으로 보는 방법

⋮

$$w_{ij} = \begin{cases} 1 & \text{if } d_{ij} < d \\ 0 & \text{otherwise} \end{cases}, \text{ where } d = \text{minimum distance}$$

- ▲ 기준 거리(d)를 너무 작게 설정하면 이웃이 없는 고립된 점이 생길 수 있으므로,
d를 각 관측치별 최단거리보다는 크게 설정해야 함!

공간가중행렬의 이웃 결정 기준

③ K-Nearest Neighbors Weights

머신러닝의 KNN 알고리즘과 비슷한 방식으로
가장 근접한 K개의 점을 이웃으로 보는 방법



이렇게 만들어진 공간가중행렬은
그대로 쓰이지 않고, 정규화하여 사용!

공간자기상관 진단

Moran's I 지수

지역 간 인접성을 나타내는 **공간가중행렬**과
 실제 인접 지역들 간의 **속성 데이터**의 **유사성**을 측정하는 방법

$$I = \frac{N \sum_i^N \sum_j^N w_{ij} (Y_i - \bar{Y})(Y_j - \bar{Y})}{(\sum_i^N \sum_j^N w_{ij}) \sum_i^N (Y_i - \bar{Y})^2}$$

$$Z_I = \frac{I - E(I)}{\sqrt{Var(I)}} \quad \text{where } E(I) = -\frac{1}{N-1}$$

N : 지역 단위 수, Y_i : i지역의 속성, Y_j : j지역의 속성, \bar{Y} : 평균값, w_{ij} : 가중치

공간자기상관 진단

Moran's I 지수

▶ I 값의 범위 : $-1 \sim 1$

▶ 검정통계량 : $Z_I \rightarrow$ **Z 검정**을 통해 전역적 공간자기상관의 유의성 판단

⋮

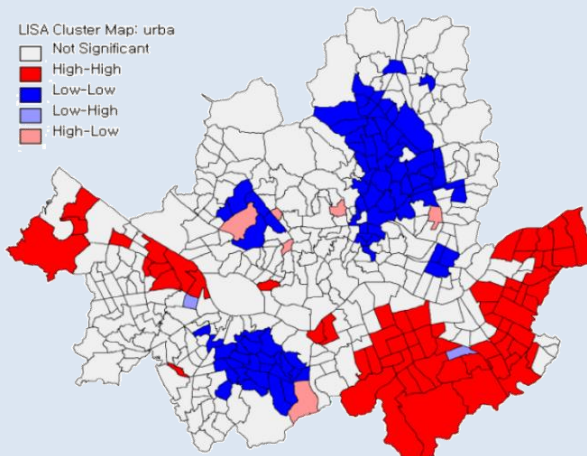
한계

전체에서 공간 자기상관이 존재하는지만 알 수 있을 뿐,
핫스팟이나 콜드스팟의 위치는 알 수 없음!

공간자기상관 진단

LISA 지표 *Local Indicator of Spatial Association*

특정 개별 지역들이 전체 지역의 공간자기상관성에
얼마나 영향을 미치는지 파악하는 **국지적** 측정 방법



공간 자기상관이 **세부적**으로
어느 지역에서 나타나는 것인지 알 수 있다!

HH(high-high), LL(low-low): 공간적 군집지역

HL(high-low), LH(low-high): 공간적 이례지역

공간자기상관 진단



공간자기상관이 **종속변수**에서 나타났는지, **오차**에서 나타났는지에 따라
사용해야 하는 공간회귀모델이 달라짐

라그랑지 승수검정 *Lagrange Multiplier*

OLS 회귀모델의 **종속변수** 또는 **오차**에서
공간자기상관이 실재하지 않는다는 귀무가설에 대해 검정하는 것

공간회귀모델 선택 과정

공간자기상관 진단

Moran's I 지수, LISA로 공간자기상관성 확인

공간자기상관이 종속변수에서 나타났는지, 오차에서 나타났는지에 따라
 사용해야 하는 공간회귀모델이 달라짐

라그랑지 승수 검정(LM-Lag, LM-Error)으로 모델 선택

라그랑지 승수검정 *Lagrange Multiplier* LM-Error

LM-Lag

	유의X	유의
유의X	기존 OLS 모델	공간오차모델
유의	공간시차모델	Robust-LM으로 다시 검정

공간자기상관 처방

앞서 보인 공간데이터의 특성(문제 유형)에 따라 해결방법이 달라짐

공간 자기상관성	공간시차모델(SLM)
	공간오차모델(SEM)
공간적 이질성	지리가중회귀모형(GWR)

공간 자기상관성 → 인접지역의 영향력을 **변수에 포함**시켜 통제

공간적 이질성 → 각 지역마다 **다른 추정계수**로 영향력을 추정

공간자기상관 처방

① 공간시차모델 (SLM, Spatial Lag Model)

인접지역의 공간적 의존성을 변수로 투입시켜서
공간시차변수를 하나의 **설명변수**로 두는 모델

$$Y = \rho WY + X\beta + \varepsilon = (1 - \rho W)^{-1}(X\beta + \varepsilon)$$

공간시차변수

EXAMPLE

주택가격 = 주택면적 + 건축년도 + 가구주의 소득 + 오차



공간시차변수 투입

주택가격 = **W*주택가격** + 주택면적 + 건축년도 + 가구주의 소득 + 오차

공간자기상관 처방

② 공간오차모델 (*SEM, Spatial Error Model*)

오차를 **공간오차변수**로 변형시킨 모델

$$Y = X\beta + \mu = X\beta + (I - \lambda W)^{-1}\varepsilon$$

공간오차

EXAMPLE

주택가격 = 주택면적 + 건축년도 + 가구주의 소득 + 오차



공간오차변수로 변형

주택가격 = 주택면적 + 건축년도 + 가구주의 소득 + **공간오차**

공간자기상관 처방

③ 지리가중회귀모델 (*GWR, Geographically Weighted Regression*)

변수들 간의 관계에 대한 **회귀계수가 지역마다 서로 다르다는 전제** 하에
지역 별로 회귀모델을 추정하는 방법

$$W_i^{1/2}Y = W_i^{1/2}X\beta_i + W_i^{1/2}X\varepsilon_i$$

$$\beta(u_i, v_i) = [X'W(u_i, v_i)X]^{-1}X'W(u_i, v_i)XY$$



회귀분석이 분석단위(지역) 별로 이루어졌기 때문에
추정된 **회귀계수** 값은 **해당 격자에서만** 의미가 있음!

공간회귀분석 흐름 정리

시각화를 통한 인사이트 도출, 공간자기상관 의심



이웃 설정 및 공간가중행렬 생성

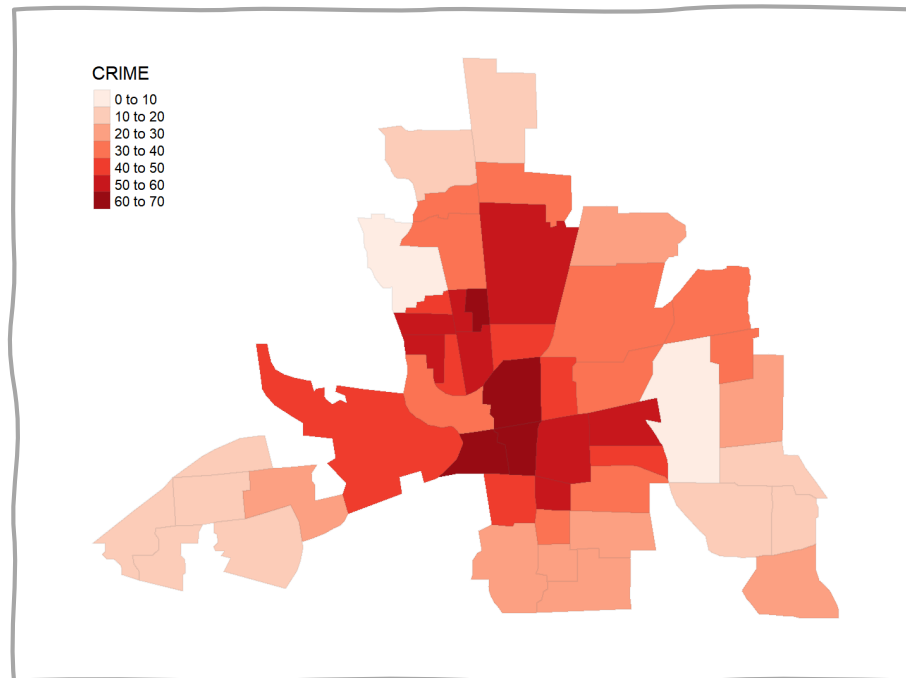


Moran's I 지수, LISA로 공간자기상관 검정

공간회귀분석 흐름 정리

시각화를 통한 인사이트 도출, 공간자기상관 의심

...



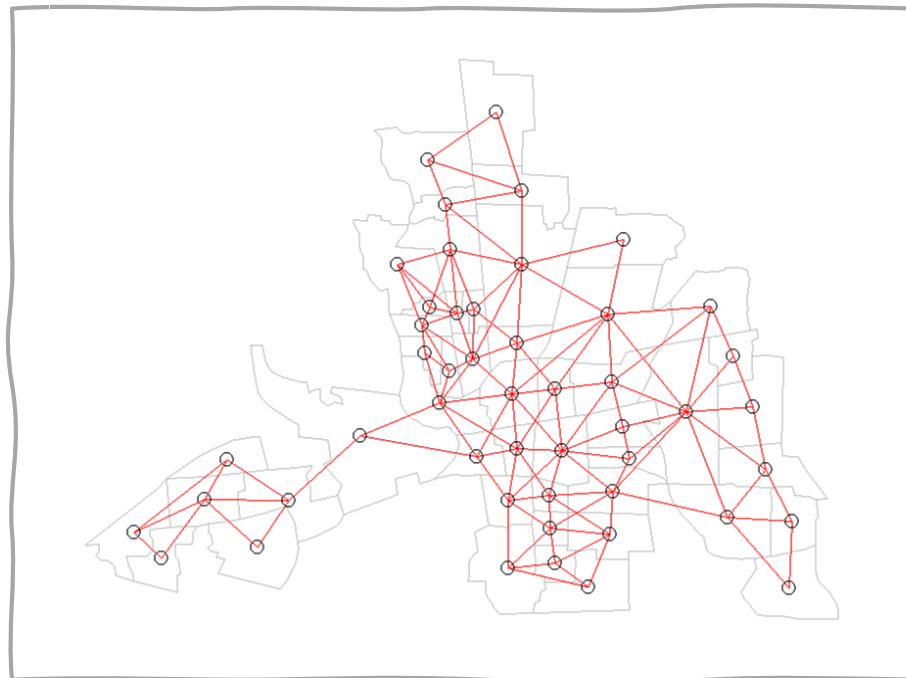
4

공간회귀분석

공간회귀분석 흐름 정리

이웃 설정 및 공간가중행렬 생성

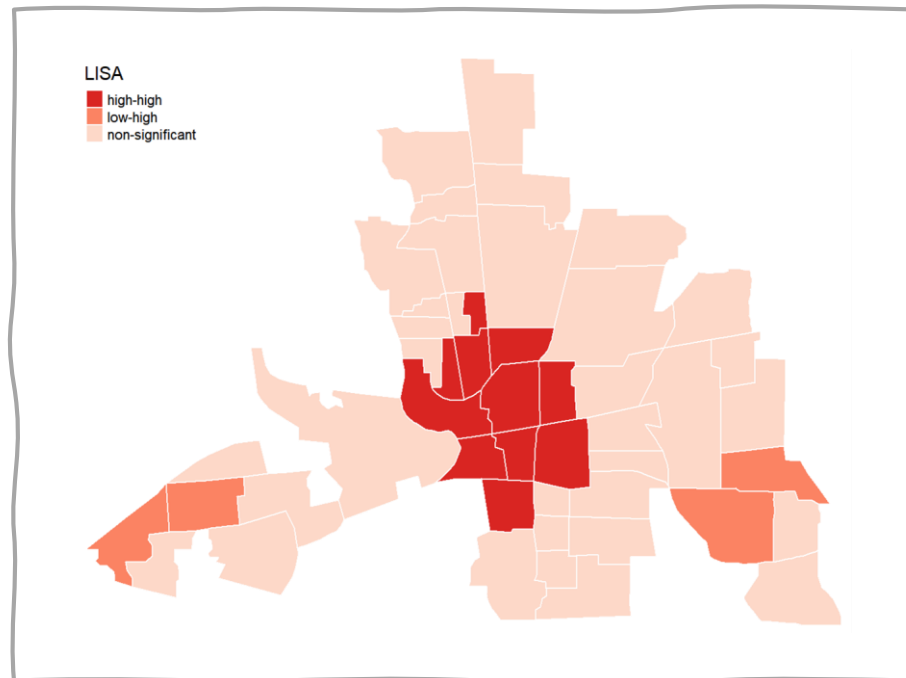
⋮



공간회귀분석 흐름 정리

Moran's I 지수, LISA로 공간자기상관 검정

...



공간회귀분석 흐름 정리

라그랑지 승수 검정으로 모델 선택



모델 적합 및 대안 모형이 기존 모형을 개선하였는지 점검

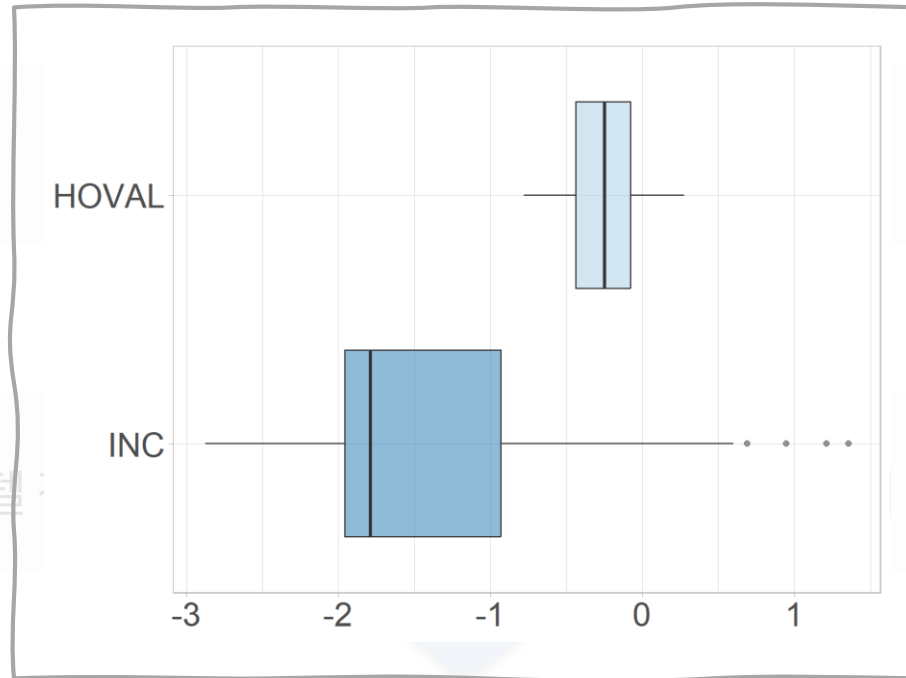


모델 해석

4

공간회귀분석

공간회귀분석 흐름 정리



모델 해석



감사합니다!

By.

단 란
회 귀



▽ 2023. 09. 21 ▽
▽ 화귀 ▽
▽ 잠진업 1등 ▽