회귀분석팀

6팀
고경현
박세령
박이현
박지성
심예진
이선민

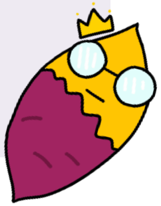
INDEX

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

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

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

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

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

추정량의 분산 커짐

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

검정통계량 감소 (t-test)

귀무가설을 기각하지 못하게 됨

1 다중공선성

해결법

다중공선성 해결방법

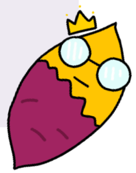
- 변수 선택
- 정규화
- 차원 축소
- 필터링

2 변수 선택법

변수 선택법이란?

수많은 변수들 중 **적절한 변수 조합**을 찾아내는 방법

서로 **상관이 있는 독립 변수**들을 **일부 제거** ⇨ **다중공선성 해결**



| 변수 선택법 |

변수가 제거되는 것에

논리성과 정당성을 부여하는 방법

변수 선택 지표

AIC Akaike Information Criterion

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

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

$\hat{\sigma}^2$: 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 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

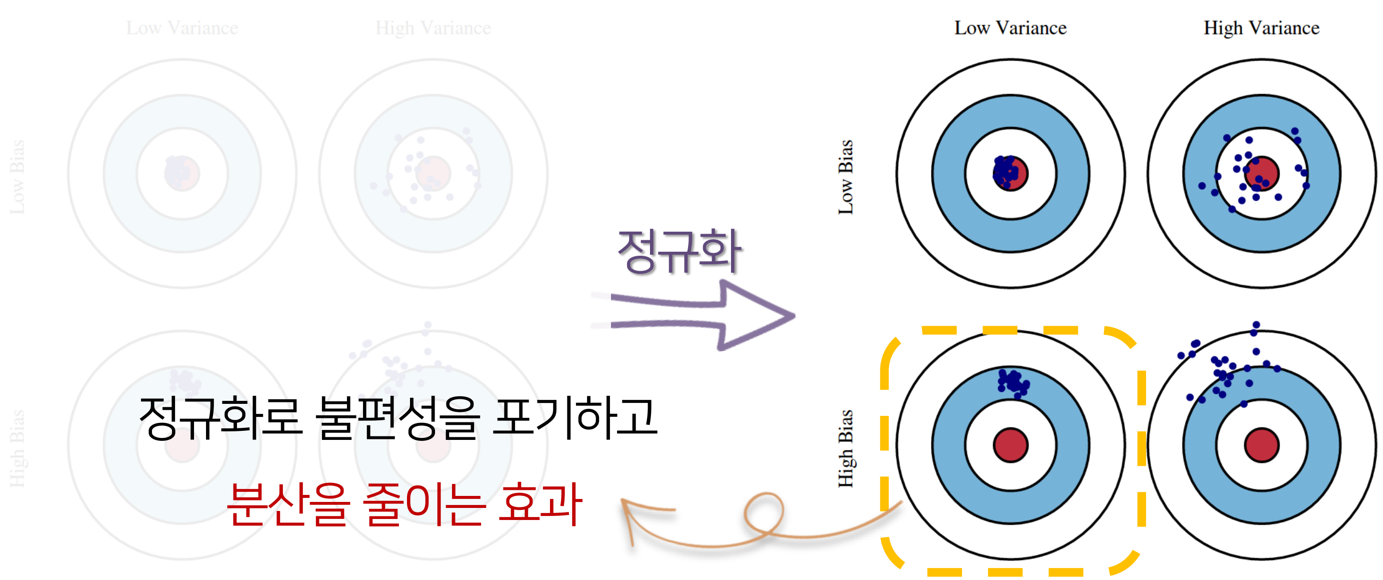
정규화

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

목적함수에 대한 이해 ②

λ의 값에 따른 회귀계수의 변화

라그랑지안 승수법을 이용해 나타낸 함수식

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

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

회귀제곱합을 최소화 & Penalty term을 통제

회귀계수의 크기 조정

Tuning Parameter

음수가 아닌 튜닝파라미터

회귀계수를 찾는 과정에서 직접 CV를 통해 조정해주는 모수

제약조건의 크기를 결정 (s와는 반대 관계)

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

Lasso Regression

Lasso Regression

SSE를 최소화하면서 회귀계수 β에 제약 조건을 거는 방법

(L1 Regularization)

제약 조건식이 L1-norm 형태

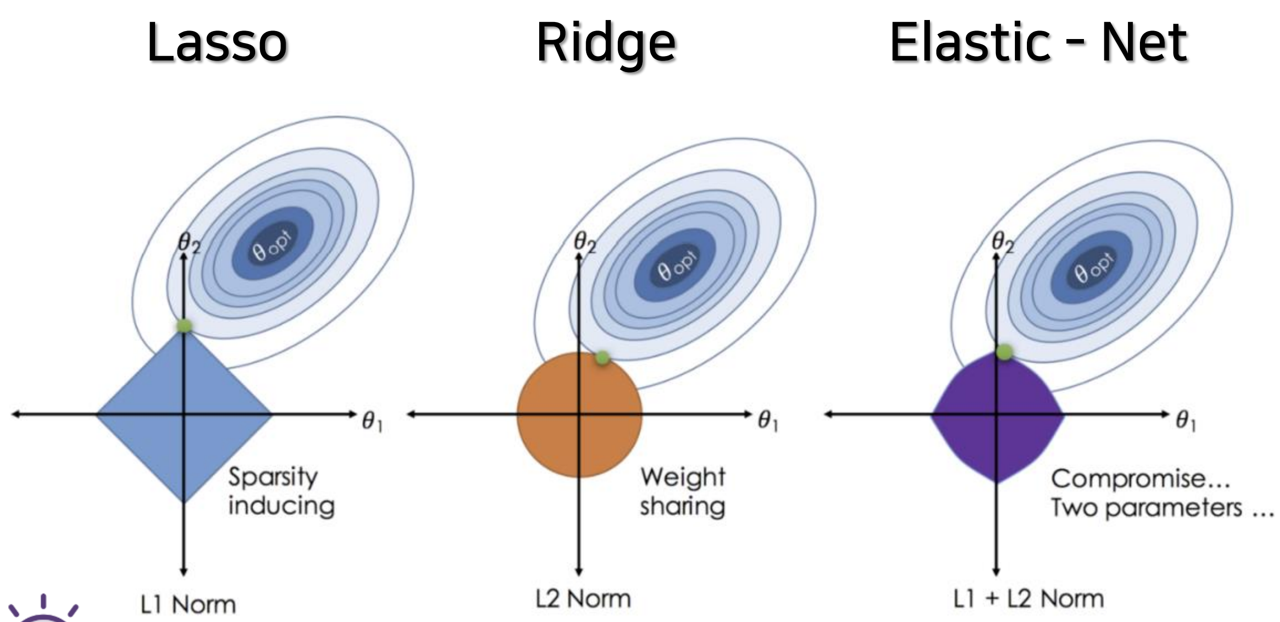
목적함수

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

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

식을 최소화하여 회귀계수의 Lasso estimator를 얻을 수 있음

Elastic-Net



제약 조건에 따라 추정량이 만들어지는 공간도 변화

Fused Lasso

Fused Lasso

변수들 사이의 물리적인 거리가 존재한다는 정보를 활용하는 모델

목적함수

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

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

Lasso

\+

인접한 변수들의 회귀 계수를 비슷한 값으로 추정하게 만드는 regularization term