회귀분석팀

6팀

심은주
진수정
문병철
이수정
임주은

INDEX

1. 다중공선성
2. 변수선택법
3. 축소 추정

0

지난 주 복습

- 회귀 가정

모델의 선형성	오차의 등분산성
오차의 정규성	오차의 독립성

모델의 선형성

- residuals vs fitted plot
- crPlots

→

- 변수 변환
- Polynomial Regression

오차의 등분산성

- Scale-location plot
- Shapiro-Wilk test

→

- Box-Cox transformation
- Yeo-Johnson transformation

오차의 정규성

- Normal QQ plot
- BP test
- Jarque-Bera test

→

- Box-Cox transformation
- Yeo-Johnson transformation

오차의 독립성

- Durbin Watson Test

→

- 시계열 분석
- 공간회귀분석

- **공간데이터의 특성: 공간자기상관**

전역적 공간자기상관
Global Spatial Autocorrelation

전체 구역이 가지는
하나의 공간자기상관의 정도

국지적 공간자기상관
Global Spatial Autocorrelation

개별 지점이 가지는
공간자기상관의 정도

- **공간데이터의 특성: 공간적 이질성**

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

- **공간가중행렬**

지역 내 다수의 지점들이 서로 공간적으로 인접하고 있는지의 여부를 파악할 수 있도록 행렬로 나타낸 것

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

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

- 공간회귀 선택 알고리즘

모란 I 지수, LISA로 공간 자기상관성이 있는지 확인

라그랑지승수검정으로 모델 선택

유의 X — OLS 회귀모델

LM-Lag 유의 — 공간시차모델

LM-Error 유의 — 공간오차모델

둘 다 유의 — Robust LM

- **공간시차모델(SLM, Spatial Lag Model)**

한 지역의 관측치가 인접지역의 관측치와 상관성이 있는 경우,

공간적 의존성을 하나의 설명변수로 둔 모델

공간시차변수

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

$$\varepsilon \sim MVN(0, \sigma^2 I_n)$$

- **공간오차모델(SEM, Spatial Error Model)**

오차에 공간자기상관성이 있는 경우, 오차를 공간오차변수로 변형시켜 준 모델

공간오차변수

$$Y = X\beta + \mu \qquad where \;\; \mu = \lambda W\mu + \varepsilon \;\; and$$

$$= X\beta + (1 - \lambda W)^{-1}, \qquad \varepsilon \sim MVN(0, \sigma^2 I_n)$$

- **지리가중회귀모델(GWR, Geographically Weighted Regression)**

변수들 간의 관계를 추정하는 회귀계수가 지역마다 서로 다른 것을 전제로 지역별로 회귀모델을 추정하는 방법

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

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

공간의 특성을 반영하고 있는 변수가 거리에 따라 얼마나 민감하게 변하는지 보여주는 지표

Exponential 가중치

$$W_i = \sqrt{\exp(-d_i/\theta)}$$

d_i: i 지역에서부터 다른 지역까지의 거리,

θ: 대역폭

1

다중공선성

- **다중공선성이란?**

예측변수 X들 간의 선형관계가 존재하는 경우

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$

나머지 X 변수들이 고정되었을 때,
x_1이 1단위 증가하면 y는 평균적으로 β_1만큼 증가함을 의미

개별 변수 해석 시, '다른 변수를 **고정**한 상태에서 해당 X의 증분'

Uncorrelated한 경우만 가능

- **다중공선성이란?**

예측변수 X들 간의 선형관계가 존재하는 경우

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$

나머지 X 변수들이 고정되었을 때,
x_1이 1단위 증가하면 y는 평균적으로 β_1만큼 증가함을 의미

개별 변수 해석 시, '다른 변수를 **고정**한 상태에서 해당 X의 증분'

Uncorrelated한 경우만 가능

정확한 회귀분석을 위해선 다중공선성이 크면 안된다!

- **다중공선성이란?**

예측변수 X들 간의 선형관계가 존재하는 경우

Y : 학점　X1: 결석 횟수　X2: 출석률　X3: 강의 수

$$출석률 = 1 - \frac{결석횟수}{강의수} \quad \rightarrow \quad X2 = 1 - \frac{X1}{X3}$$

X2를 X1과 X3의 선형결합으로 완벽하게 설명 → X2는 필요 없는 변수!

- **행렬로 이해**

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \dots & \dots & & \dots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix}$$

완전한 다중공선성이 존재하면,

$Y = X\beta + \epsilon$ 에서 행렬 X의 rank가 **full rank가 아님**

↓

$X'X$ 의 **역행렬 존재 X** $(Det(X) = 0)$

↓

$\hat{\beta} = (X'X)^{-1}(X'Y)$ 도 구할 수 없음

- **행렬로 이해**

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \dots & \dots & & \dots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix}$$

완전한 다중공선성이 아니더라도…

$$\hat{\beta} = (X'X)^{-1}(X'Y), \quad Var(\hat{\beta}) = \sigma^2 (X'X)^{-1}, \quad (X'X)^{-1} = \frac{1}{Det(X'X)} adj(X'X)$$

$Det(X'X)$ 이 0에 가까워질수록,

$(X'X)^{-1}$ 이 커져 분산 역시 커지고,

회귀계수 추정이 매우 불안정

- **다중공선성의 문제점**

회귀계수들의 분산이 커져 t 검정 통계량이 작아짐

귀무가설 ($\beta = 0$) 을 기각하지 못하는 경우 발생

전체 회귀식은 유의하지만,

개별 회귀계수 중에는 유의한 것이 없는 결과 발생

왜 이런 일이 발생하나요…?

"특정 변수 x_j가 이미 고정된 다른 변수x_k에 의해 설명되므로,

x_k가 x_j의 몫까지 설명해버려서 개별 회귀계수는 유의 X"

결과적으로, Prediction Accuracy가 심각하게 감소!

- **다중공선성의 판별법**

직관적인 판단

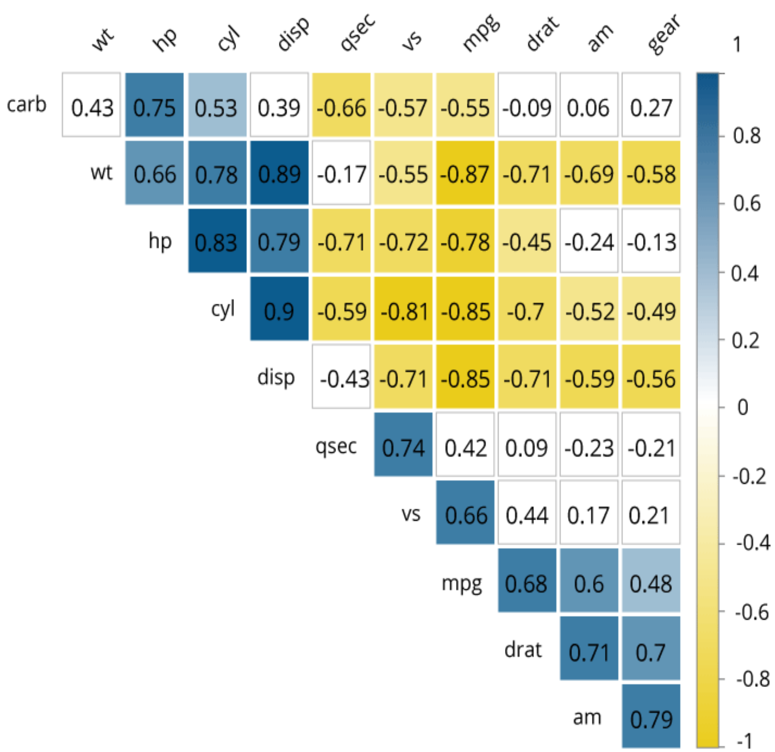
- F-test는 유의하나 개별 회귀계수들에 대한 t-test에서 귀무가설을 대부분 기각하지 못하는 경우
- 상식적으로 유의한 회귀계수가 유의하지 않다고 나올 경우
- 회귀계수의 부호가 상식과 다를 경우

WHY?

다른 변수가 이미 해당 변수의 영향력을 설명하고 있기 때문!

- **다중공선성의 판별법**

상관계수 플랏



- 절댓값 기준
  상관계수가 0.7이상일 경우
  다중공선성 의심

- R의 ‘Corrplot’ 패키지 이용

- **다중공선성의 판별법**

VIF(Variance Inflation Factor, 분산팽창인자)

$$VIF = \frac{1}{1-{R^2}_j}, \quad j = 1, \dots, p$$

${R^2}_j$: x_j를 $x_1 + \cdots + x_p$ 으로 회귀식을 적합했을 때, 도출되는 R^2 값

${R^2}_j$가 높다

x_j가 다른 변수들로 충분히 설명될 수 있다

다중공선성 존재

- **다중공선성의 판별법**

VIF(Variance Inflation Factor, 분산팽창인자)

$$VIF = \frac{1}{1-R^2{}_j}, \quad j = 1, \dots, p$$

$R^2{}_j$: x_j를 $x_1 + \cdots + x_p$ 으로 회귀식을 적합했을 때, 도출되는 R^2 값

- VIF가 1이면, 다중공선성 없음 ($R^2{}_j$=0이므로)

- VIF가 10이상이면, 심각한 다중공선성 존재

- 다중공선성의 판별법

다중공선성, 해결할 순 없을까…?

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

변수선택법	차원축소	축소추정

R_j^2 : x_j를 $x_1 + \cdots + x_p$ 으로 회귀식을 적합했을 때, 도출되는 R^2 값

둘만 다뤄볼까…?

- VIF가 1이면 다중공선성 없음 ($R_j^2=0$이므로)
- VIF가 10이상이면, 심각한 다중공선성 존재

차원축소는
SUNDAE에게 맡긴다!

2

변수선택법

- **변수선택법이란?**

후보 변수들 중에서 불필요한 변수들을 제거하여 적절한 변수들의 집합을 찾는 방법

다중공선성이 존재할 때 많이 사용!

그래서 뭐가 좋은데..?

- 높은 상관관계를 가지는 변수들 중 일부만을 선택 가능
- 높은 상관관계를 가지는 변수들의 존재를 정당화

변수선택법을 통해 최종 모델에 대한 확신을 얻을 수 있음!

- **변수선택의 척도**

Partial F-test를 통한 변수 선택

1주차 내용 기억해보자…!

H_0: $\beta_{q+1} = \beta_{q+2} = \cdots = \beta_p = 0$ RM이 적절

H_1: $\beta_{q+1}, \beta_{q+2}, \ldots, \beta_p$ 중 적어도 하나는 0이 아니다 FM이 적절

팀장의 사심*^^*

검정통계량

$$F = \frac{\frac{SSR(FM) - SSR(RM)}{p-q}}{\frac{SSE(FM)}{n-p-1}} \sim F_{p-q,n-p-1}$$

- **변수선택의 척도**

Partial F-test를 통한 변수 선택

Partial F-test

$$SSR(FM) - SSR(RM)$$

$$=SSR(X_1, \dots, X_p) - SSR(X_1, \dots, X_q)$$

$$= SSR(X_{q+1}, X_{q+2} \dots, X_p \mid X_1, X_2, \dots, X_q)$$

X_1부터 X_q까지의 변수로 회귀식을 설명하고 있는 상태에서

X_{q+1}부터 X_p까지의 변수가 추가되었을 때의 설명력

즉, SSR 값이 작으면 $X_{q+1} \sim X_p$ 변수들이 무의미함을 의미해

삭제할 수 있다!

- **변수선택의 척도**

Partial F-test를 통한 변수 선택

" 비교하려는 두 모델이 Nested되어 있을 때만 사용 가능 "

Model A ⊂ Model B 처럼 변수들 집합의 포함관계 성립

	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:	$y = \beta_0 + \beta_1 x_1 + \beta_4 x_4$	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

그럼 Nested가 아닌 경우에는…?

FM과 RM을 비교하는 Partial F-test로는 비교 불가능

Global한 모델 간의 비교를 가능하게 해주는 기준이 필요!

뒤에서 알아보자…!

- **변수선택의 척도**

수정결정계수(R_a^2)

벌써 까먹은 건 아니쥐?

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

회귀식의 설명력을 의미

수정결정계수가 더 큰 모델을 사용하자!

- **변수선택의 척도**

Mallows(C_p)

$$C_p = \frac{SSE_p}{\hat{\sigma}^2} + (2p - n)$$

p : RM에서 사용한 독립변수 개수

$\hat{\sigma}^2$: FM의 오차항의 분산의 추정값

n : 관측치의 개수

- 변수를 추가하면 SSE는 무조건 작아지기 때문에, 2p를 패널티로 넣음

- $C_p \approx p$ 일수록 bias가 작은, 좋은 모델

- C_p값을 이용하여 모델을 비교할 때에는 동일한 독립변수의 전체 집합을 가진 모델일 경우에만 사용 가능

- **변수선택의 척도**

AIC(Akaike Information Criterion)

$$AIC_p = n \ln\left(\frac{SSE_p}{\hat{\sigma}^2}\right) + 2p$$

모델의 적합도

패널티

적합도와 변수의 개수를 동시에 고려

적절한 복잡도를 가진 변수 조합 완성

AIC 값이 작을수록 더 좋은 모델!

- **변수선택의 척도**

BIC(Bayesian Information Criterion)

$$BIC_p = n \ln\left(\frac{SSE_p}{\hat{\sigma}^2}\right) + p \ln n$$

$n > 8$ 이면 AIC보다 변수 개수에 더 많은 패널티 부여

변수 개수가 더 적은 모델 선택

BIC 값이 작을수록 더 좋은 모델!

- **변수선택의 종류**

변수선택법은 경험적인 방법

명확한 답이 존재하는 것이 아니라 ,

직접 알고리즘에 따라 해당하는 모든 경우를 계산해서 가장 좋은 회귀식을 찾는다!

Best Subset Selection	전진선택법	후진제거법	단계적 선택법

- Best Subset Selection (All Possible Regression)

가능한 모든 변수들의 조합 고려

1. M_1부터 M_p까지 구하기
   - $M_k(k = 1, \dots, p)$란 변수의 개수를 k개로 적합했을 때 적합한 회귀식 중 MSE가 제일 작은 식
2. 위에서 배운 척도를 이용해 M_1부터 M_p중 최적의 회귀식 찾기
   - 보통 Mallows C_p나 BIC 사용

- Best Subset Selection (All Possible Regression)

가능한 모든 변수들의 조합 고려

1. M_1부터 M_p까지 구하기
   - $M_k(k = 1, \dots, p)$란 변수의 개수를 k개로 적합했을 때 적합한 회귀식 중 MSE가 제일 작은 식
2. 위에서 배운 척도를 이용해 M_1부터 M_p중 최적의 회귀식 찾기
   - 보통 Mallows C_p나 BIC 사용

장점

- 모든 경우를 다 고려하기 때문에 Best Model에 대한 신뢰도 ↑

단점

- $p > 40$인 경우 계산이 불가
- 많은 관측치를 지니고 있다면 계산 비용이 큼

- **전진선택법 (Forward Selection)**

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

1. Null Model($y = \beta_0$)에서 시작
2. X_1부터 X_p까지의 변수 중에 AIC와 BIC를 가장 크게 낮추는 변수 추가
3. 만약 2번에서 X_1이 선택되었다면, $y = \beta_0 + \beta_1 x_1$의 식에 X_2부터 X_p까지의 변수 중에 AIC와 BIC를 가장 크게 낮추는 변수 추가
4. 위의 과정 반복
   - AIC와 BIC가 낮아지면 변수 추가
   - AIC와 BIC가 낮아지지 않으면 과정 중단

- **전진선택법 (Forward Selection)**

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

장점

- 계산량이 Best subset selection에 비해 적음

단점

- 변수를 추가하는 과정에서 모든 조합을 고려하지 않음

→ 최적의 모델이라고 할 수 없음!

- 후진제거법 (Backward Elimination)

Full Model($y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$) 에서 시작해 변수를 하나씩 제거하는 방법

1. Full Model($y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$)에서 시작해 모든 변수들 중에서 제거했을 때 가장 AIC와 BIC를 크게 낮추는 변수를 선택해 제거

2. 위의 과정 반복
   - AIC와 BIC가 낮아지면 변수 제거
   - AIC와 BIC가 낮아지지 않으면 과정 중단

- **후진제거법 (Backward Elimination)**

Full Model($y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$) 에서 시작해 변수를 하나씩 제거하는 방법

장점

- 계산량이 Best subset selection에 비해 적음

단점

- 데이터의 개수(n)< 변수의 개수(p)일 때 사용 불가
- 변수를 추가하는 과정에서 모든 조합을 고려하지 않음

→ 최적의 모델이라고 할 수 없음!

- **단계적 선택법 (Stepwise Selection)**

Forward Selection과 Backward Elimination 과정을 섞은 방법

1. 먼저 forward selection 과정을 이용해 가장 유의한 변수들을 모델에 추가
2. 나머지 변수들에 대해 Backward Elimination을 적용해 새롭게 유의하지 않게 된 변수들 제거
3. 위의 과정 반복
   - 제거된 변수는 다시 모형에 포함되지 않음
   - 추가했을 때 유의한 설명변수가 더 이상 없을 때까지 반복

- **단계적 선택법 (Stepwise Selection)**

Forward Selection과 Backward Elimination 과정을 섞은 방법

장점

- 계산량이 Best subset selection에 비해 적음
- 변수를 선택할 수도 제거할 수도 있기 때문에 더 유연하게 움직임

단점

- 변수를 추가하는 과정에서 모든 조합을 고려하지 않음

→ 최적의 모델이라고 할 수 없음!

변수선택법 문제점!

경험적인 방법이기에 계산량이 굉장히 많음

특히, Best Subset Selection은 계산량이 정~~말 많음

Forward Selection, Backward Elimination, Stepwise Selection은

모든 경우의 수를 고려하지 않기 때문에

기계적으로 변수를 제거하는 것은 위험

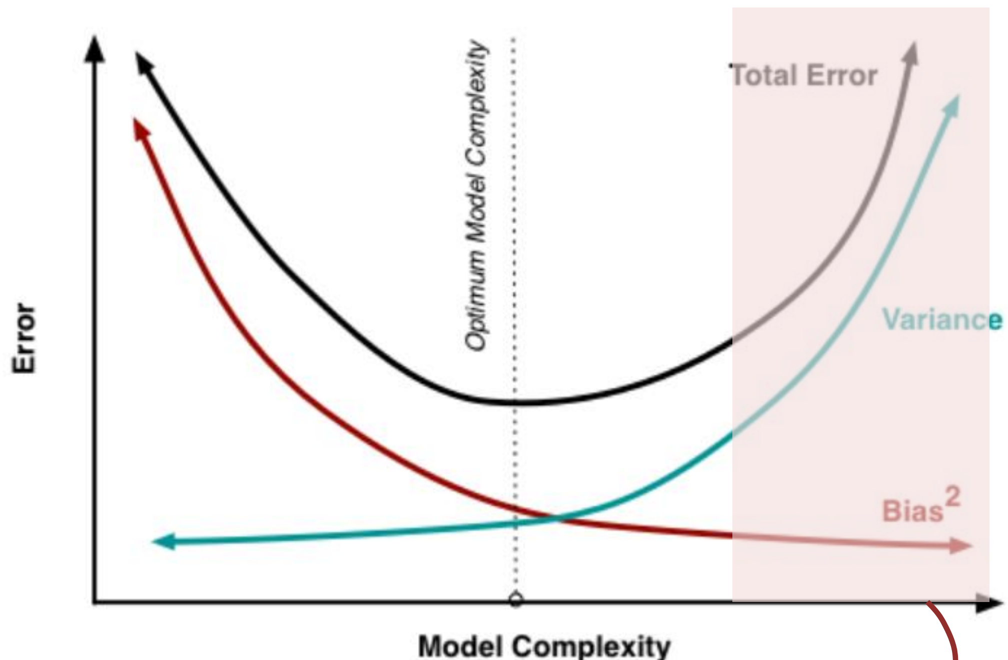
축소 추정 방법 추천!!

3

축소 추정

- **축소 추정이란?**

각각 개별 베타 추정량을 **0으로 수축**시키는 방법

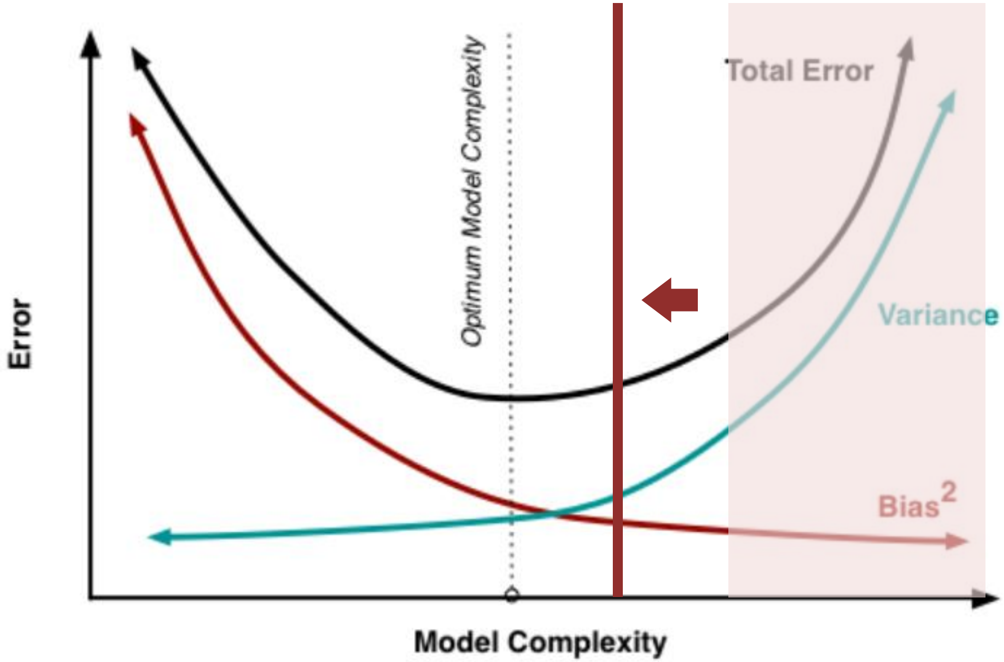


그렇다면
추정량의 일부 편향을 허용하되,
분산을 더 줄일 수 있다면 어떨까?

다중공선성이 존재할 경우, 각각 개별 베타 계수의 분산이 매우 크게 상승

- **축소 추정이란?**

각각 개별 베타 추정량을 **0으로 수축**시키는 방법



둘 다 가질 순 없지…

불편성을 포기하되,

전체 MSE($Bias^2 + Variance$)를 더 작게 하는

추정량 얻을 수 있다!

- Ridge Regression (L2 Regularization)

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

- 위의 식을 최소화하는 β값을 가짐
- 앞 부분: SSE, 회귀식이 데이터에 잘 적합하여 SSE를 작게 만드는 계수 추정치를 찾음
- 뒷 부분: shrinkage penalty로 β값들이 0에 가까울 수록 작아짐

→ **이 항이 계수 추정치들을 0으로 축소하는 영향을 가짐**

- Ridge Regression (L2 Regularization)

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

조절 모수 λ: SSE항과 패널티항을 조절하는 역할

- λ=0 이면, 패널티 효과가 없어 최소제곱추정치 생성
- $\lambda \to \infty$ 이면, 패널티 효과가 커져 계수 추정치가 0에 가까워짐

∴ Ridge regression은 각각의 **λ 값에 따라서 다른 추정치** 집합들을 만듦

- Ridge Regression (L2 Regularization)

$$\hat{\beta}^{ridge} = \underset{\beta}{argmin} \sum_i (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

조절 모수 λ: SSE항과 패널티항을 조절하는 역할

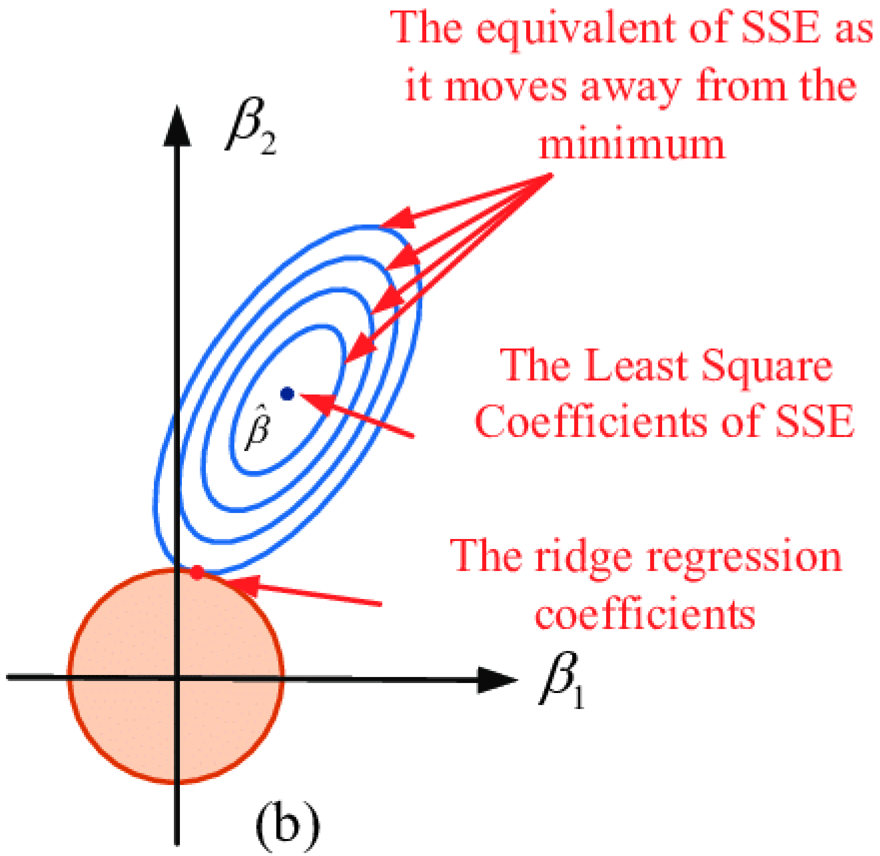
좋은 λ 값 선택 중요

주로 CV를 통해 튜닝

- λ=0 이면, 패널티 효과가 없어 최소제곱추정치 생성
- $\lambda \rightarrow \infty$ 이면, 패널티 효과가 커져 계수 추정치가 0에 가까워짐

∴ Ridge regression은 각각의 λ 값에 따라서 다른 추정치 집합들을 만듦

- Ridge Regression (L2 Regularization)



1

베타의 개수가 2개인

Ridge regression의 제약범위는

원으로 표현 가능 $(\beta_1^2 + \beta_2^2 \leq s)$

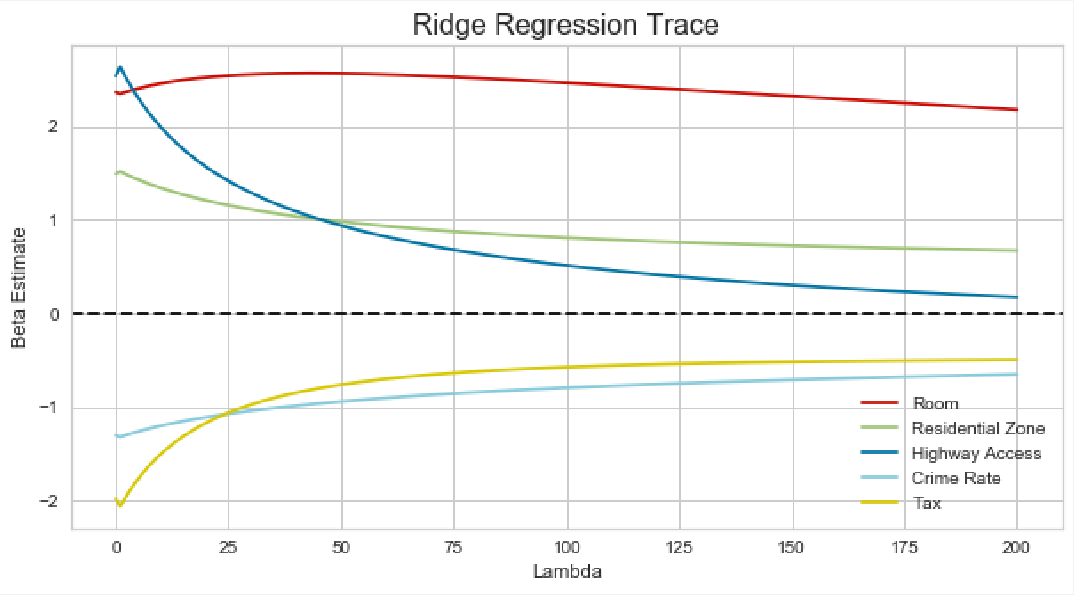
2

회귀계수가 0에 가까워 지긴 하지만

λ 가 ∞ 가 아닌 이상 0이 될 수 없음

- Ridge Regression (L2 Regularization)

Ridge regularization path



- 변수 제거하지 못함
- 예측에는 상관이 없지만, 해석할 때 문제 발생 가능

Ridge regression을 통해 만들어진 회귀계수는 0에 가까울 뿐 0은 아님

- Ridge Regression(L2 Regularization)

Ridge regularization path

회귀계수를 0으로 만들어

변수를 제거할 수 있는 축소추정법은 없을까?

- 변수 제거하지 못함
- 예측에는 상관이 없지만, 해석할 때 문제 발생 가능

Ridge regression을 통해 만들어진 회귀계수는 0에 가까울 뿐

Lotso 아니고 Lasso 등장

- Lasso Regression (L1 Regularization)

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

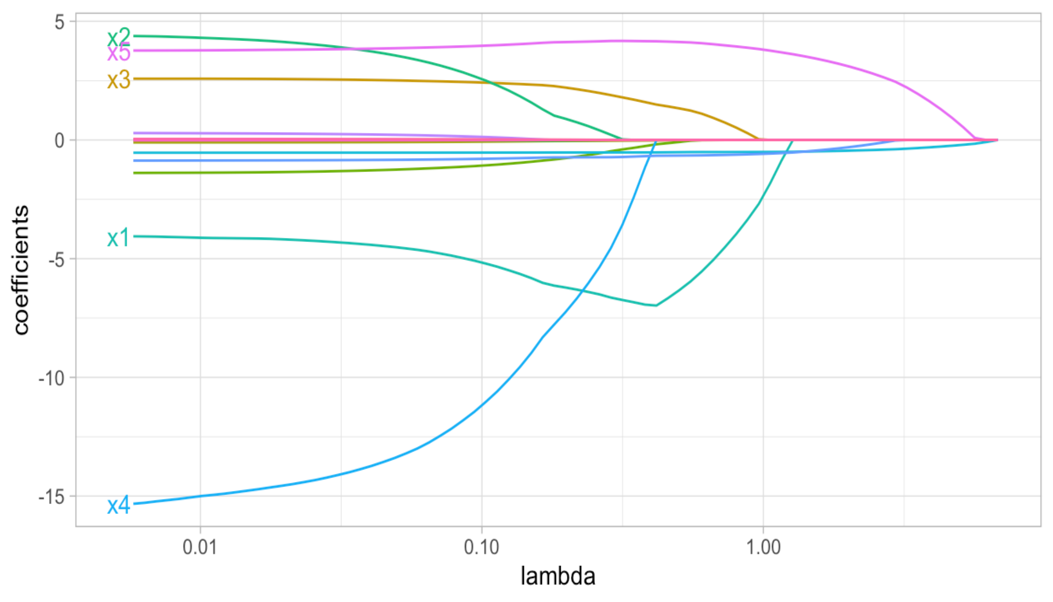
Ridge와 비슷하게 패널티를 주어 계수 추정치들을 0으로 축소하는 방법

- Penalty term이 절댓값으로 들어감
- Ridge와 달리 변수 선택의 효과가 있음

변수 잘 가고~~

- Lasso Regression (L1 Regularization)

Lasso regularization path

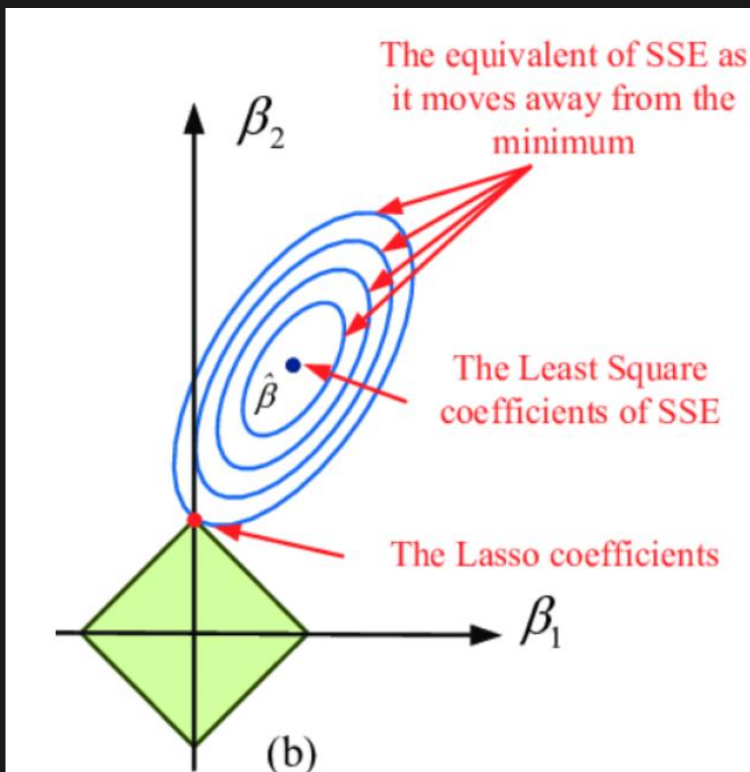


- λ가 커지면, 추정계수는 0으로 축소됨
- 변수 제거 가능

Lasso regression을 통해 만들어진 회귀계수는 0이 될 수 있음!

왜 변수선택이 가능한지는 다음 슬라이드에서!

Lasso이 변수 선택의 효과를 갖는 이유는?



Penalty term이 **절댓값**이므로 **제약범위**가 날카로운 **모서리**를 가지는 형태

↓

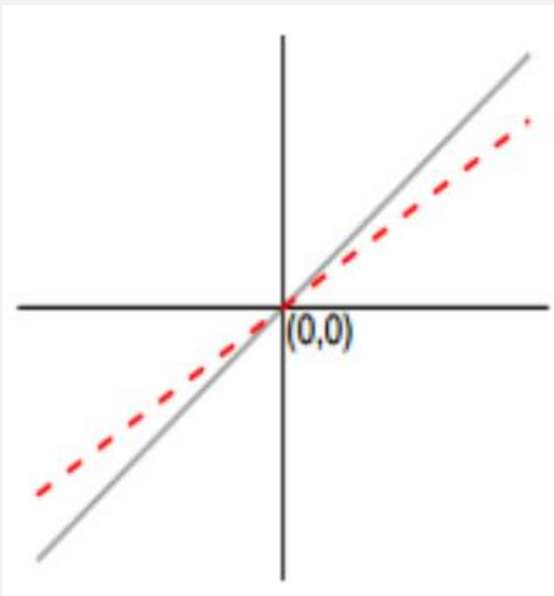
최적값이 **모서리** 부분에서 나타날 확률이 높음

↓

몇몇 계수를 **0으로 추정**, 즉 **변수 선택**!

- Ridge vs. Lasso

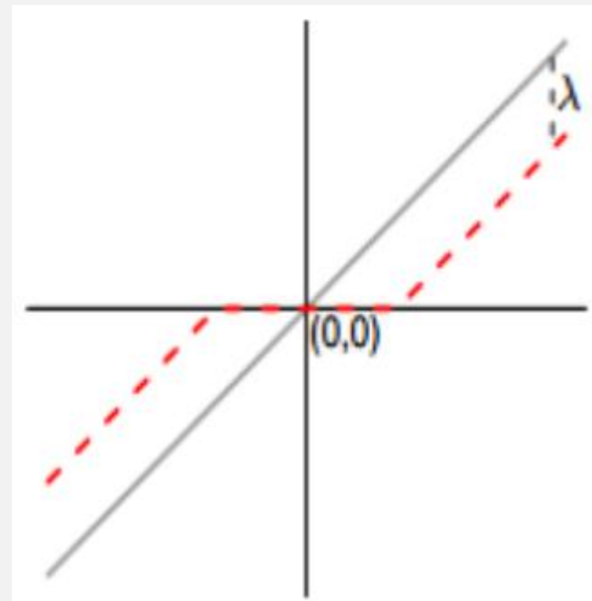
Ridge regression



추정계수의 크기가
같은 비율로 축소됨

VS.

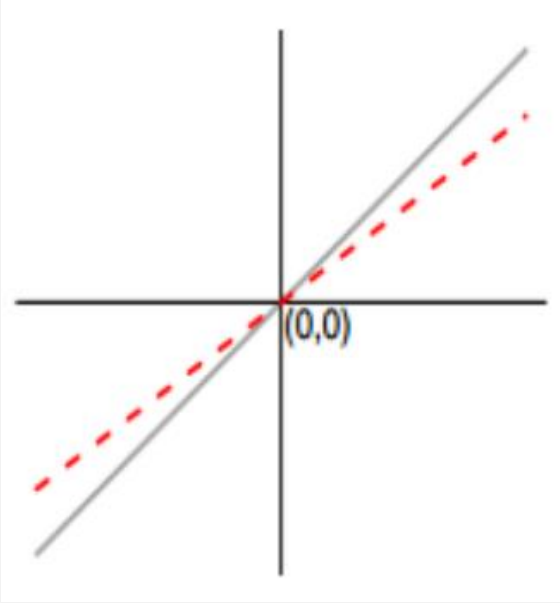
Lasso regression



추정계수의 크기가
같은 양만큼 축소됨

- Ridge vs. Lasso

Ridge regression

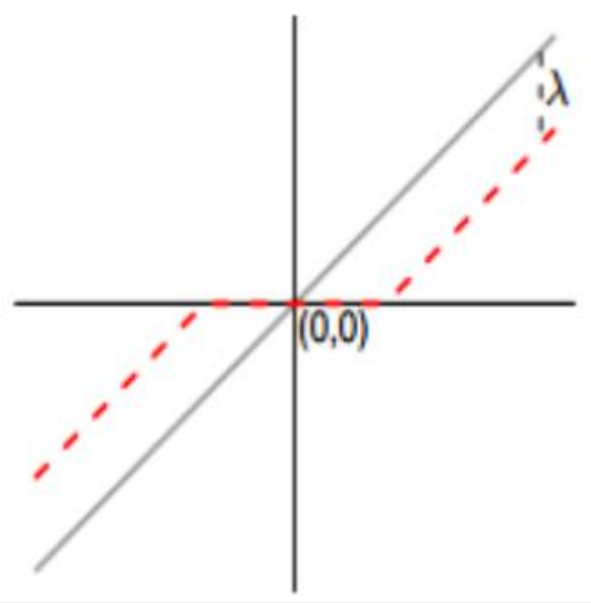


상관성이 있는 변수들에 대해서

적절한 가중치 배분

VS.

Lasso regression



상관성이 있는 변수들에 대해

추정 계수의 크기가 작은 변수를 제거

- **Elastic Net**

Lasso와 Ridge를 절충하여 각각의 단점을 보완한 정규화 방법

Ridge part Lasso part

$$L_{enet}(\hat{\beta}) = \frac{\sum_i (y_i - x_i^T \beta)^2}{2n} + \lambda\left(\left(\frac{1-\alpha}{2}\right)\sum_j \beta_j^2 + \alpha \sum_j |\beta_j|\right)$$

α 는 Ridge와 Lasso의 가중치로,

$\alpha = 0$이면 식은 Ridge가 되고, $\alpha = 1$이면 Lasso가 됨

- Adaptive Lasso Regression

Oracle Properties

1. 올바른 변수들로 이루어진 부분 집합 모델을 식별한다.
2. 최적 추정 속도를 갖는다.

→ 최적의 추정 속도로 정확한 변수 선택이 가능한가?

좋은 추정 절차는 Oracle Properties를 충족해야 함

- Adaptive Lasso Regression

But ***Lasso regression…***

- λ 값에 따라 변수 선택이 consistent하지 않음
- 자동적인 변수 선택이 가능하지만, 크게 편향된 추정치를 반환

→ **언제나 Oracle Properties를 만족하는 것은 아님!**

A-haaaa !!!

각각의 β 계수에 따른 weight로 패널티를 주어

Bias를 줄이자는 것이 Adaptive Lasso

- Adaptive Lasso Regression

$$\widehat{\beta^*}^{(n)} = \underset{\beta}{argmin} \left\| y - \sum_j x_j \beta_j \right\|^2 + \lambda_n \sum_j \widehat{w}_j |\beta_j|$$

$$where \ \widehat{w}_j = \frac{1}{|\hat{\beta}|^{\gamma}}$$

$\hat{\beta}$: initial estimate (ex. OLS)

- 계수 추정량이 클수록 작은 가중치를 줌으로써, Sparsity는 유지하되 Bias를 줄일 수 있음

- λ를 적절히 선택한다면, Oracle Properties를 만족하게 됨

- Adaptive Lasso Regression

$$\widehat{\beta^{*}}^{(n)} = \underset{\beta}{argmin} \left\| y - \sum_j x_j \beta_j \right\|^2 + \lambda_n \sum_j \widehat{w}_j |\beta_j|$$

$$where\ \widehat{w}_j = \frac{1}{|\hat{\beta}|^{\gamma}}$$

$\hat{\beta}$: initial estimate (ex. OLS)

가중치 W는

어떻게 설정할까?

- 계수 추정량이 클수록 작은 가중치를 줌으로써, Sparsity는 유
  Bias를 줄일 수 있음

- λ를 적절히 선택한다면, Oracle Properties를 만족하게 됨

- **가중치 w를 설정하는 방법**

Two-stage approach

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

$where\ \ \hat{\beta}_j : initial\ estimate$

λ 와는 관계없이

가중치는 $\hat{\beta}_j$ 에 의해 결정됨

Path-wise approach

$$w_j(\lambda) = w\big(\hat{\beta}_j(\lambda)\big)$$

$where\ \ \hat{\beta}_j : initial\ estimate$

가중치 값이

λ 의 변화에 영향을 받게 됨

자세한 설명은 생략한다…

Adaptive Lasso 계산 알고리즘에 적용되는 아이디어

$$\widehat{\beta^*}^{(n)} = \underset{\beta}{argmin} \left\| y - \sum_j \frac{x_j}{\hat{w}_j}\hat{w}_j\beta_j \right\|^2 + \lambda_n \sum_j \hat{w}_j|\beta_j|$$

$\tilde{\beta}_j = \hat{w}_j\beta_j$ 로 치환

$$\widehat{\beta^*}^{(n)} = \underset{\beta}{argmin} \left\| y - \sum_j \frac{x_j}{\hat{w}_j}\tilde{\beta}_j \right\|^2 + \lambda_n \sum_j \tilde{\beta}_j$$

$x_j^{**} = x_j/\hat{w}_j$ 라고 보면, 결국 Lasso를 푸는 문제로 변형됨!

구해진 $\tilde{\beta}_j$ 에 대해 $\tilde{\beta}_j/\hat{w}_j$ 를 계산하여 최종적인 $\widehat{\beta^*}^{(n)}$ 값 도출

이제 이것을 적용해서 계산해보자!

- Adaptive Lasso 계산 알고리즘

Step1

초기 추정치 $\hat{\beta}$ 추정

다중공선성 X → OLS 추정치 사용

다중공선성 O → Ridge 추정치 사용

Step2

가중치 설정 방식에 따라 $\hat{\beta},\ \lambda_n,\ \gamma$ 를 적절히 이용하여 $\hat{w}$ 계산

이 때, $\lambda_n,\ \gamma$ 는 CV를 통해 정함

- Adaptive Lasso 계산 알고리즘

Step3

$x_j^{**} = x_j / \widehat{w}_j$ 라고 정의

Step4

모든 λ_n 에 대해 Lasso problem을 풂

$$\hat{\beta}^{**} = \underset{\beta}{argmin} \left\| y - \sum_j x_j^{**} \beta_j \right\|^2 + \lambda_n \sum_j |\beta_j|$$

Step5

Adaptive Lasso의 추정치 $\widehat{\beta^*}_j^{(n)} = \hat{\beta}_j^{**} / \widehat{w}_j$

THANK YOU