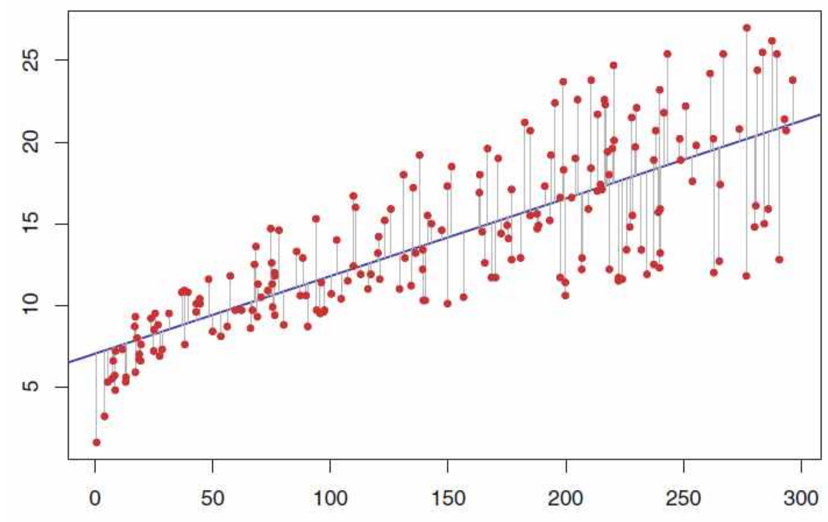
Linear Regression



: 종속 변수 Y와 한 개 이상의 독립 변수 X와의 선형 상관 관계를 모델링하는 기법

X = [1, 2, 3] -> Y = [3, 5, 7]

Question) X=4일 때, Y=?

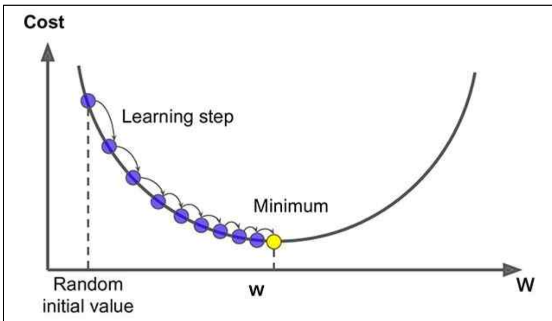
Answer) f(x) = 2x+1

ex) H(W,b) = Wx+b(목표: W=2, b=1)

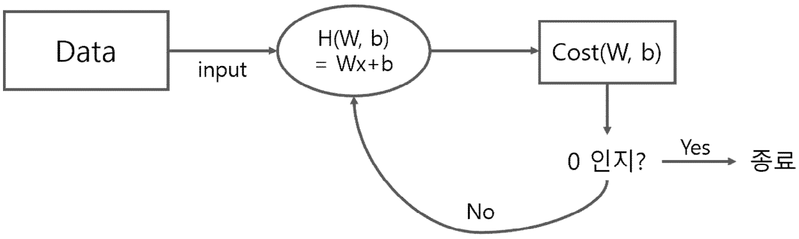
[가설 초기화]

(W=1, b=0) -> 얼마나 잘못되었는가? **Cost(W, b) -> 최소제곱법**

$Cost(W,b) = \frac{1}{m} \times \sum_{i=1}^{m} (y_i - Wx_i - b)^2$ - 최소제곱법(OLS) $\min_{W,b} \sum_{i=1}^{m} (y_i - Wx_i - b)^2$	- 제곱을 사용하는 이유 1) 제곱을 이용하면 비용이 더 커진다 → 가설이 잘못 되었을 때 그에 대한 패널티를 더 강하게 주어 빠르게 학습 2) 절대값을 이용하면 연산속도 감소

	Gradient Descent : Cost를 줄이기 위해 반복적으로 기울기를 계산하여 변수의 값을 변경해나가는 과정 기울기가 음수라면 오른쪽으로 이동 (+) 양수라면 왼쪽으로 이동 (-)

- 경사하강법



업데이트 : 1000↑(epoch)

$W = W - \alpha \frac{\partial \, Cost(W,b)}{\partial W}$

$b = b - \alpha \frac{\partial \, Cost(W,b)}{\partial b}$

정규화 모델 -> 제약 부여

선형회귀 계수에 대한 제약조건을 추가함으로써 모형이 과도하게 최적화되는 현상 방지

테스트 데이터에 대한 예측 성능

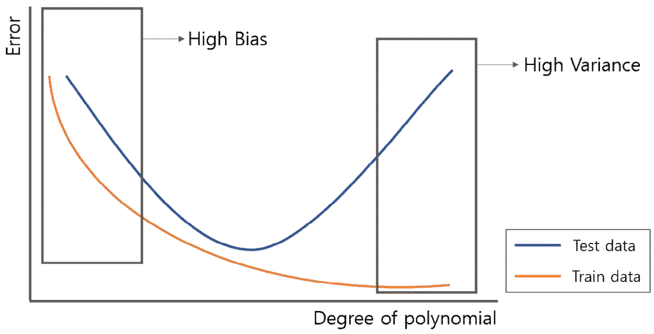
: Expected MSE = $Irreducible\,Error + Bias^2 + Variance$

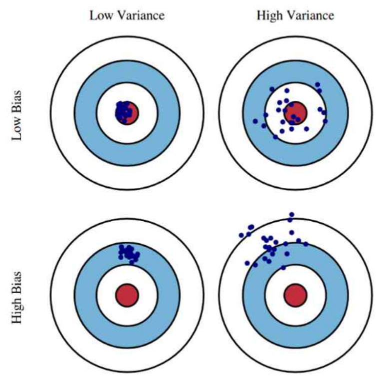
- Subset Selection

: 전체 p개의 설명변수(X) 중 일부 k개만을 사용하여 회귀계수 beta 추정

→ 전체 변수 중 일부만 선택하면 bias가 증가하지만, variance 감소

- Best subset selection
- Forward stepwise selection
- Backward stepwise elimination
- Least angle regression
- Orthogonal matching pursuit

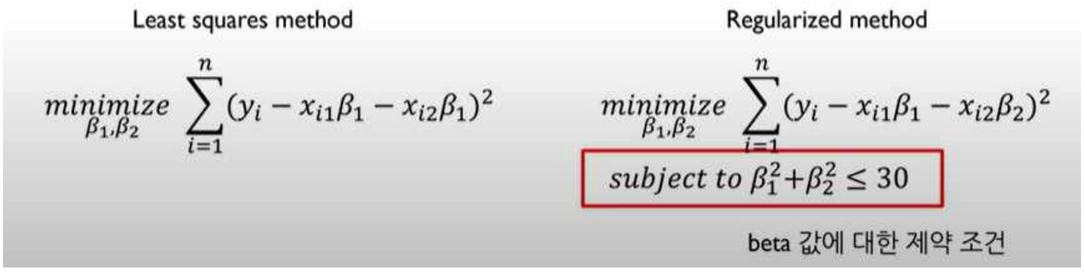




$L(\beta) = \min_{\beta} \underbrace{\sum_{i=1} (y_i - \hat{y}_i)^2}_{\text{(1) Training accuracy}} + \underbrace{\lambda \sum_{j=1}^{p} \beta_j^2}_{\text{(2) Generalization accuracy}}$	(2) Generalization accuracy를 추가하면서 베타에 제약을 주어 정규화 가능 λ: regularization parameter that controls the tradeoff between (1) and (2) · λ very big → $\beta_i \approx 0$ → high bias → underfitting · λ very small → high variance → overfitting

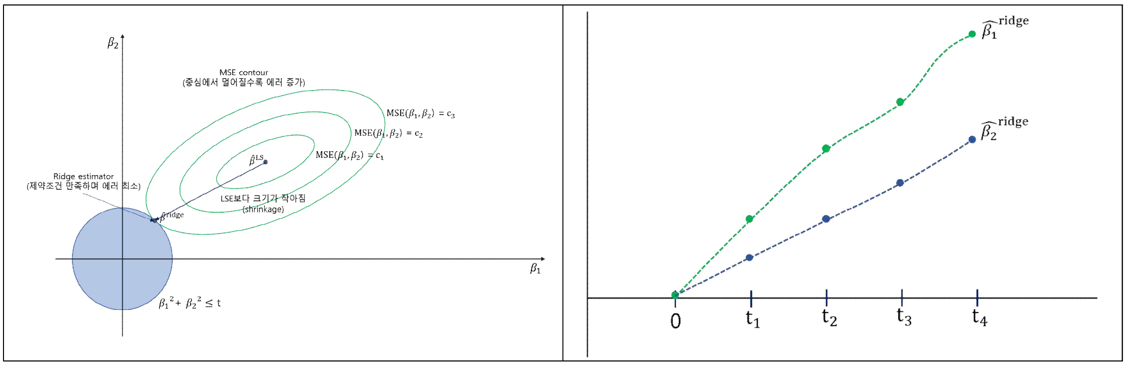
Regularization method는 회귀 계수 beta가 가질 수 있는 값에 제약 조건 부여
: 제약조건에 의해 **bias 증가, variance 감소**

ex)

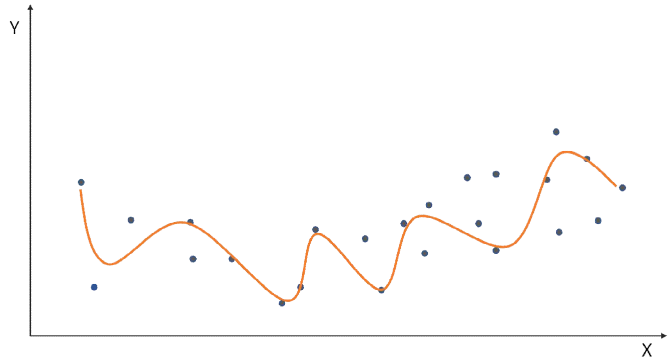


1\. Ridge Regression
제곱 오차를 최소화하면서 회귀계수 β의 L_2-norm 제한

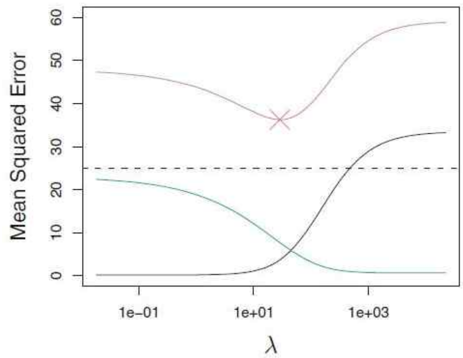
$$\hat{\beta}^{ridge} = \text{argmin}\sum_{i=1}^{n}(y_i - x_i\beta)^2,\ \text{subject to } \sum_{j=1}^{p}\beta_j^2 \le t = \text{argmin}\ \{\ MSE + \lambda\sum_{j=1}^{p}\beta_j^2\ \}$$



과대적합이 된 경우, 그래프가 극단적으로 오르락내리락 하며 선형 회귀 계수 매우 커진다.
variance를 줄이기 위해 ridge regression 사용
β^2 사용하기 때문에, 어떤 계수가 덜 중요하더라도 완전히 0으로 수렴하지 않고 충분히 작은 소수점으로 남아 있음



*Ridge regression은 변수의 크기가 결과에 큰 영향을 미치기 때문에, 변수 scaling 필요

	λ가 작을수록 계수 크기에 영향을 덜 미치므로 좀 더 flexible한 모델 형성 → 검은선인 bias 문제가 적고 초록선인 variance 문제 발생

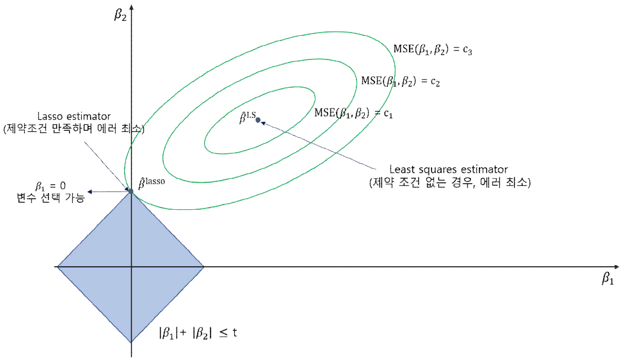
2. Lasso(Least Absolute Shrinkage and Selection Operator) Regression

: 회귀계수 β의 $L_1 - norm$ 제한, 변수 선택 가능, 꽤 robust한 편

$$\hat{\beta}^{lasso} = \text{argmin} \sum_{i=1}^{n} (y_i - x_i\beta)^2, \text{ subject to } \sum_{j=1}^{p} |\beta_j| \le t = \text{argmin} \{ MSE + \lambda \sum_{j=1}^{p} |\beta_j| \}$$

t↓, λ↑: 제약을 많이 가한다.

t↑, λ↓: 제약을 거의 가하지 않는다.



*회귀계수가 0인 변수는 y값을 예측하는데 중요하지 않은 변수, 0이 아닌 변수는 중요한 변수

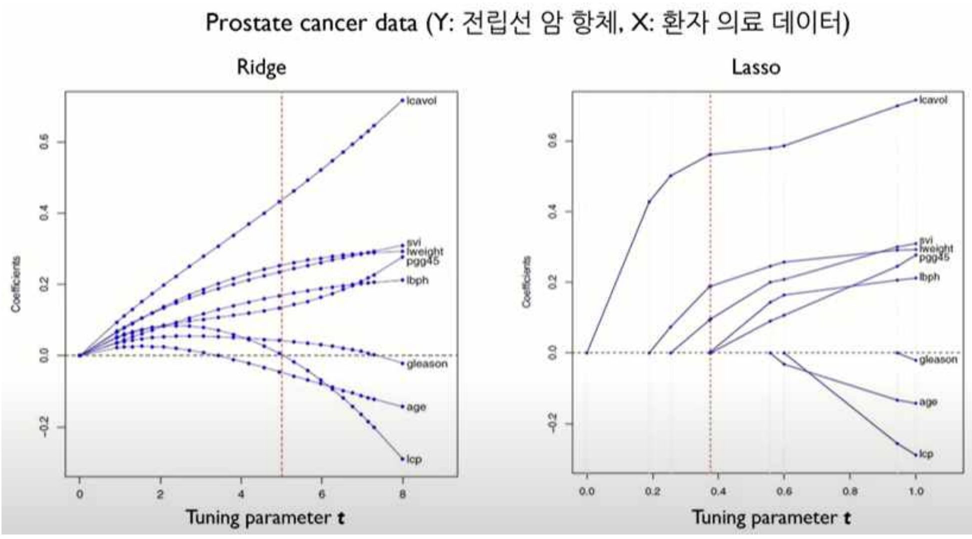
Ridge와 달리 Lasso 함수는 미분 불가능하기 때문에 회귀계수를 행렬식이 아닌, Numerical optimization methods(수치 최적화)를 이용하여 구한다.

*Numerical optimization methods

- Quadratic programming techniques(1996, Tibshirani)
- LARS algorithm(2004, Efron et al.)
- Coordinate descent algorithm(2007, Friedman et al.)

$\hat{\beta}^{lasso} = \underset{\beta}{\text{argmin}} \left\{ \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda \sum_{j=1}^{p} \lvert \beta_j \rvert \right\}$	· λ = ∞ → 회귀계수들이 0이 되어 예측값은 상수 : 적은 변수, 간단한 모델, 해석 쉬움, 높은 학습 오차(underfitting 위험 증가) · λ = 0 → 최소제곱법과 동일 : 많은 변수, 복잡한 모델, 해석 어려움, 낮은 학습 오차(overfitting 위험 증가)

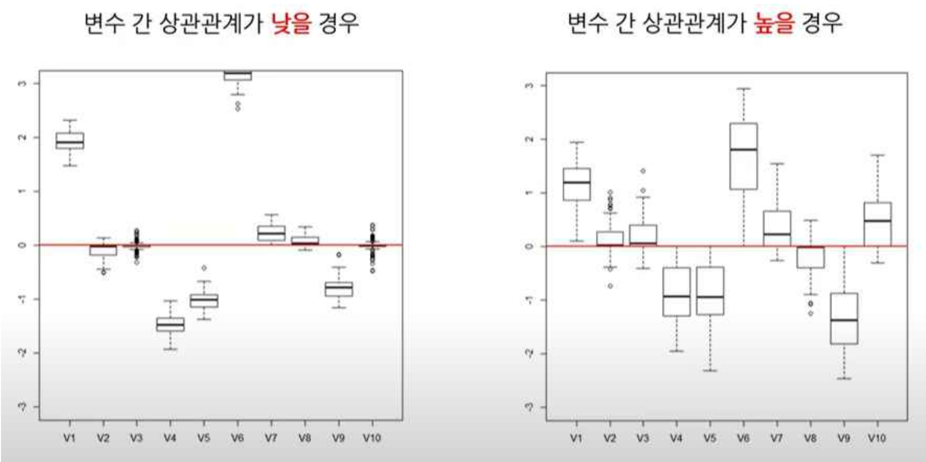
Ridge vs Lasso



- Ridge와 Lasso 모두 t가 작아짐에 따라 모든 계수의 크기 감소
- Lasso: 예측에 중요하지 않은 변수는 더 빠르게 감소
  t가 작아짐에 따라 예측에 중요하지 않은 변수는 0

	Ridge	Lasso
함수	$L_2 - norm$ regularization	$L_1 - norm$ regularization
변수 선택 가능여부	X	O
Closed form solution 존재 여부	O (미분으로 구함)	X (numerical optimization 이용)
변수 간 상관관계가 높은 상황	좋은 예측 성능	ridge에 비해 상대적으로 예측 성능↓
	크기가 큰 변수 우선적으로 줄이는 경향 O	

- 변수 간 상관관계가 높은 상황



다른 데이터에 대해서 lasso 회귀 적합했을 때, 변화하는 회귀계수 범위 시각화

3. Elastic Net

: Elastic net = Ridge + Lasso(L_1 −andL_2 −regularization)

, 상관관계 큰 변수를 동시에 선택/배제하는 특성

$$\hat{\beta}^{enet} = \operatorname{argmin} \sum_{i=1}^{n} (y_i - x_i\beta)^2, \text{ subject to } s_1 \sum_{j=1}^{p} |\beta_j| + s_1 \sum_{j=1}^{p} \beta_j^2 \le t$$

$$= \operatorname{argmin} \left\{ MSE + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right\}$$

Elastic net estimator에 대해 다음 부등식이 성립함:

$$\boxed{\left|\hat{\beta}_i^{enet} - \hat{\beta}_j^{enet}\right| \le \frac{\sum_{i=1}^{n}|y_i|}{\lambda_2}\sqrt{2(1-\rho_{ij})}}$$

ρ_{ij} X_i와 X_j 상관계수

$$\rho_{ij} = 1 \Rightarrow \left|\hat{\beta}_i^{enet} - \hat{\beta}_j^{enet}\right| \le 0$$

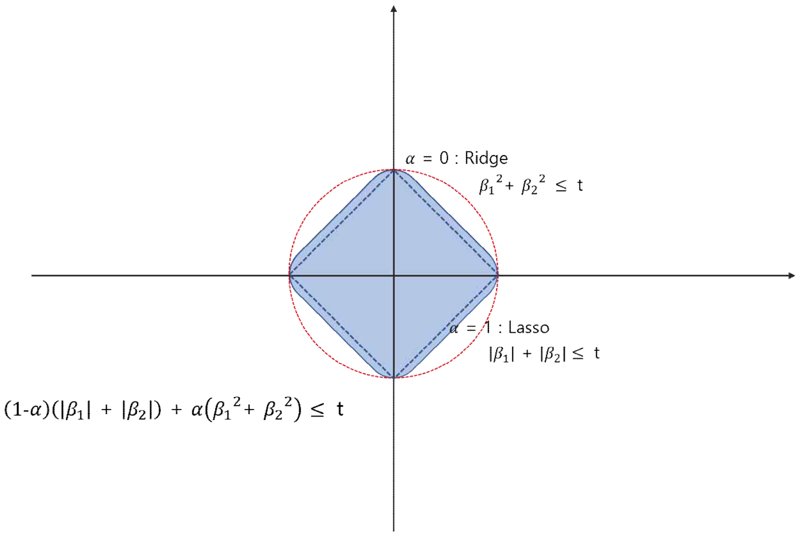
$$\Rightarrow \hat{\beta}_i^{enet} = \hat{\beta}_j^{enet}$$

$$\rho_{ij} \uparrow \text{ or } \lambda_2 \uparrow \;\Rightarrow \left|\hat{\beta}_i^{enet} - \hat{\beta}_j^{enet}\right| \downarrow$$

Grouping effect!

(Zou and Hastie, 2005)

→ (일정 범위 내에서) λ_1, λ_2 Grid Search



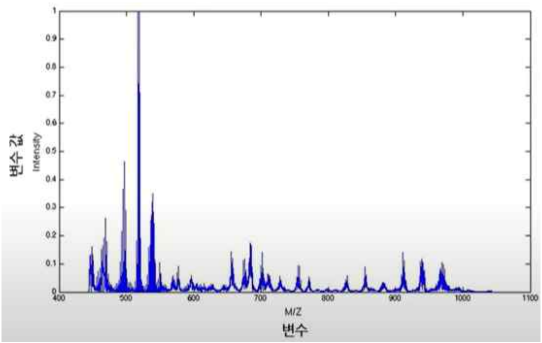
***그 외 정규화 모델**

Prior Knowledge	Regularization Method
상관관계가 높은 변수들 동시에 선택	Elastic Net
(물리적으로)인접한 변수들 동시에 선택 (상관관계와 관련 X)	Fused Lasso
사용자가 정의한 그룹 단위로 변수 선택/배제	Group Lasso
사용자가 정의한 그래프의 연결 관계에 따라 변수 선택	Grace

- Fused Lasso

: Signal/ Profile/ **Spectra** 데이터에 사용

 한 peak을 구성하는 일부 변수들이 중요하다고 했을 때, 물리적으로 그 주변에 있는 것들이 중요하다고 뽑혀야지 peak이 뽑힘



Sparsity (variable selection, L_1-norm of Lasso)

인접한 변수 beta값 비슷한 크기로 추정 (smoothness)

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

Sparsity and smoothness via the fused Lasso

(Tibshirani and Saunders, 2005)

→ 두 번째 항으로 인해 양 옆의 변수들이 차이를 최소화 하므로 동시에 뽑히게 만들도록 함

- Group Lasso

$$\hat{\beta}^{GL} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda \sum_{l=1}^{m} \sqrt{p_l} \left\| \beta^{(l)} \right\|_2 \right\}$$

$p_l =$ Number of variables in group l (group size)

$$\left\| \beta^{(l)} \right\|_2 = \left(\sum_{j=1}^{pl} \beta_{lj}^2 \right)^{\frac{1}{2}}$$

Squared sum of betas in group l

그룹 단위의 변수 선택: Group-wise sparsity

*중요하지 않아도 그룹 안에 포함되면 해당 변수 모두 선택한다는 문제점 발생

- Sparse Group Lasso

: 그룹 내에서 한 번 더 중요한 변수 선택 과정 거침

Sparse group Lasso = Lasso + group Lasso

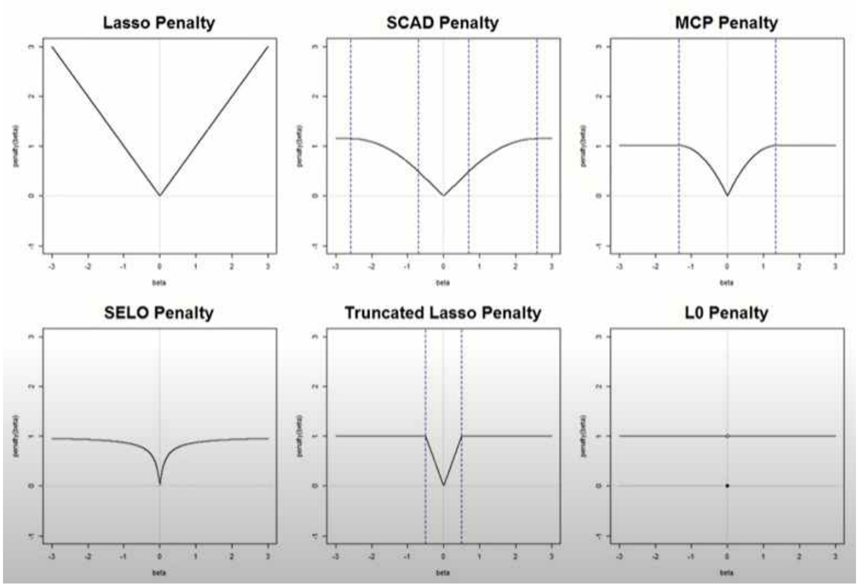
$$\hat{\beta}^{SGL} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{l=1}^{m} \sqrt{p_l} \left\| \beta^{(l)} \right\|_2 \right\}$$

그룹 단위의 변수 선택: Group-wise sparsity

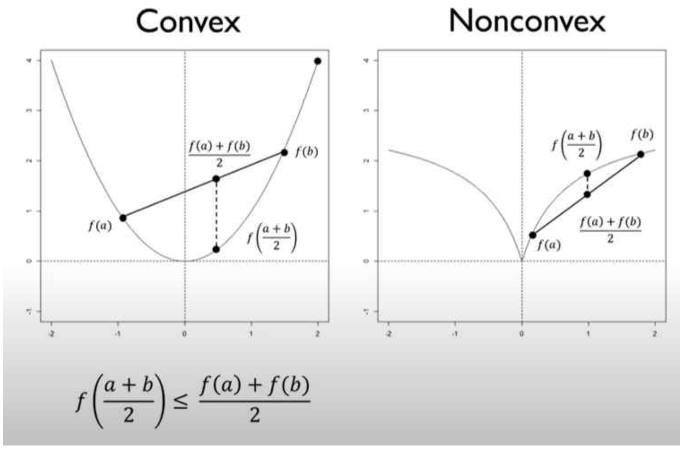
그룹 내부에서의 변수 선택: Within-group sparsity

Nonconvex Penalties

- 종류



Convex Penalties vs Nonconvex Penalties



: convex penalty는 큰 beta를 우선적으로 줄이는 효과
nonconvex penalty는 작은 beta를 우선적으로 줄이는 효과