Ⅱ. 일반 회귀분석 (Regression)

❸ Linear Regression Ⅲ

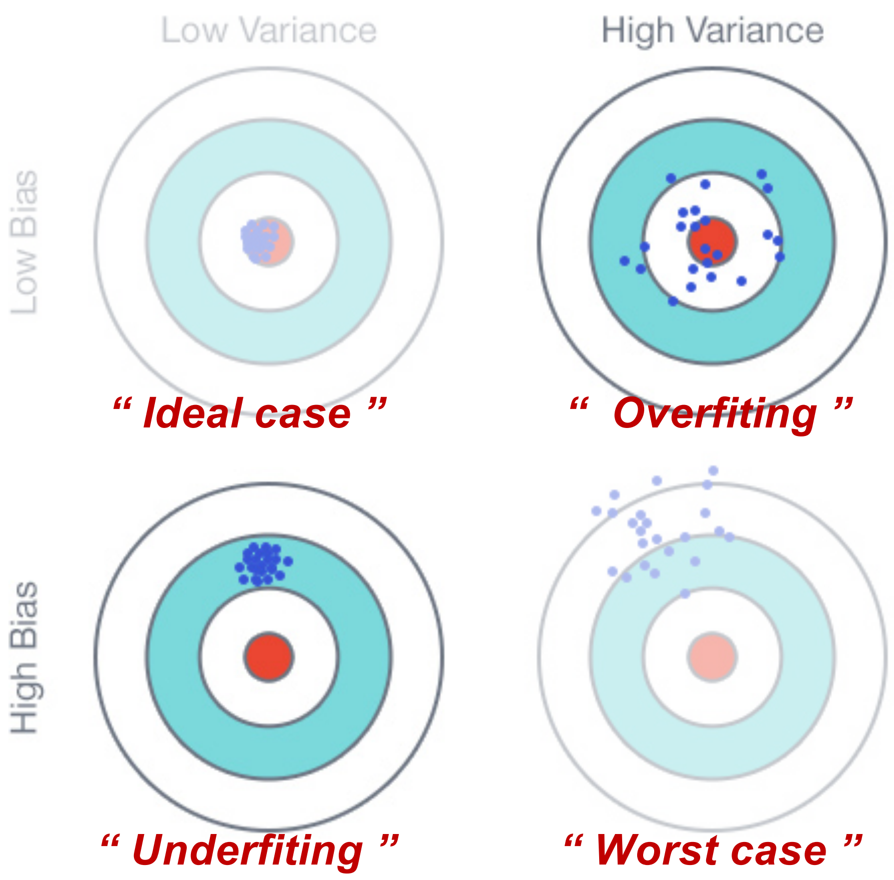
※ 참고 - 편의(Bias) vs 분산 (Variance) Trade-off

분산이 적은 모형이 좋은가, 편의가 적은 모형이 좋은가?

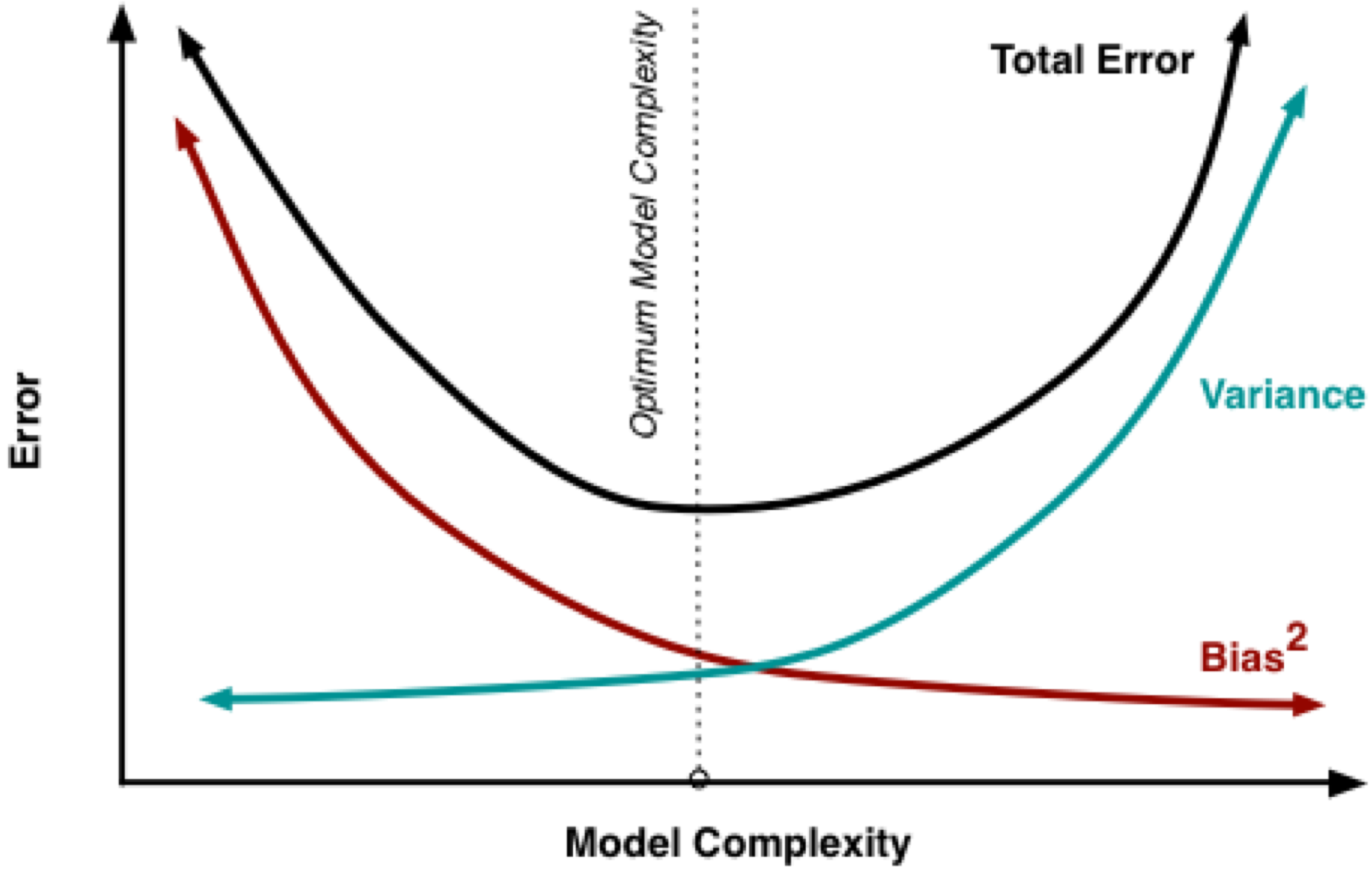
$$MSE = Variance + (Bias)^2$$

평균제곱오차(Mean Square Error; MSE)는 분산(Variance)와 편의(Bias)로 분해할 수 있으며, 평균제곱오차가 동일할 경우, 분산과 편의 간에는 트레이드 오프(Trade-off) 관계가 있음

따라서, 분산과 편의를 모두 줄이는 가장 이상적인 경우(Ideal case)가 아니라면, 분산과 편의의 적절한 수준에서 최적의 모형을 찾는 것이 가장 바람직함



※ 참고 - 편의(Bias) vs 분산 (Variance) Trade-off



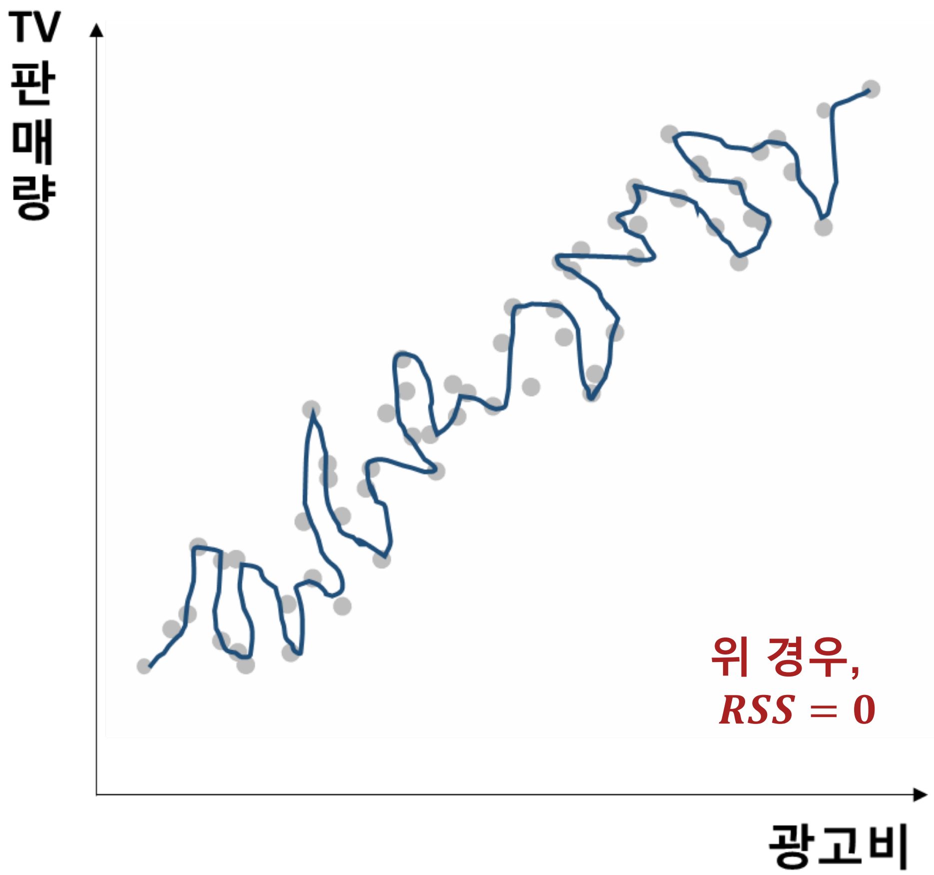
잔차제곱합 (RSS :Residual sum of squares)

MSE(Mean Square Error)가 낮으면 무조건 좋은 모형일까?

- RSS를 이용해 MSE를 계산하고, 이를 기준으로 MSE가 가장 낮은 모형을 택함으로써 변수선택에 활용하는 것이 전통적인 시각
- 하지만, 과적합(Overfitting)일 경우 RSS는 0에 근접하게 되어 올바른 변수선택을 할 수 없음

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y})^2$$

- 변수를 무한히 넣으면 모형의 설명력은 무한히 증가함



전통적인 변수선택 방법

1. Mallow's Cp 통계량('70년대)

$$C_p = \frac{RSS_d}{\sigma^2} + 2d - n$$

2. AIC(Akaike Information Criterion)

$$AIC = \frac{RSS_d}{\sigma^2} + 2d$$

3. BIC(Bayesian Information Criterion)

$$BIC = \frac{RSS_d}{\sigma^2} + 2d * log(n)$$

L1, L2 벌점화 최소제곱법

$$\hat{\beta}^{ols} = \arg\min ||Y - X\beta||^2$$

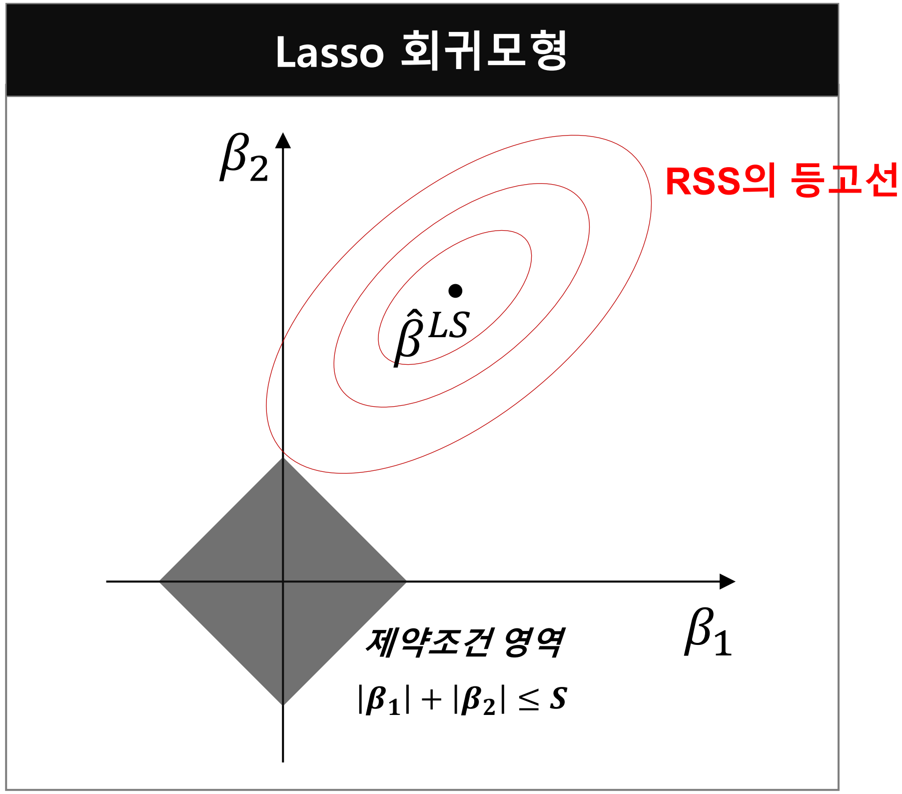
$$\hat{\beta}^{ridge} = \arg\min ||Y - X\beta||^2 - \lambda||\beta||^2 = \left(X'X + \lambda I_p\right)^{-1} X'Y$$

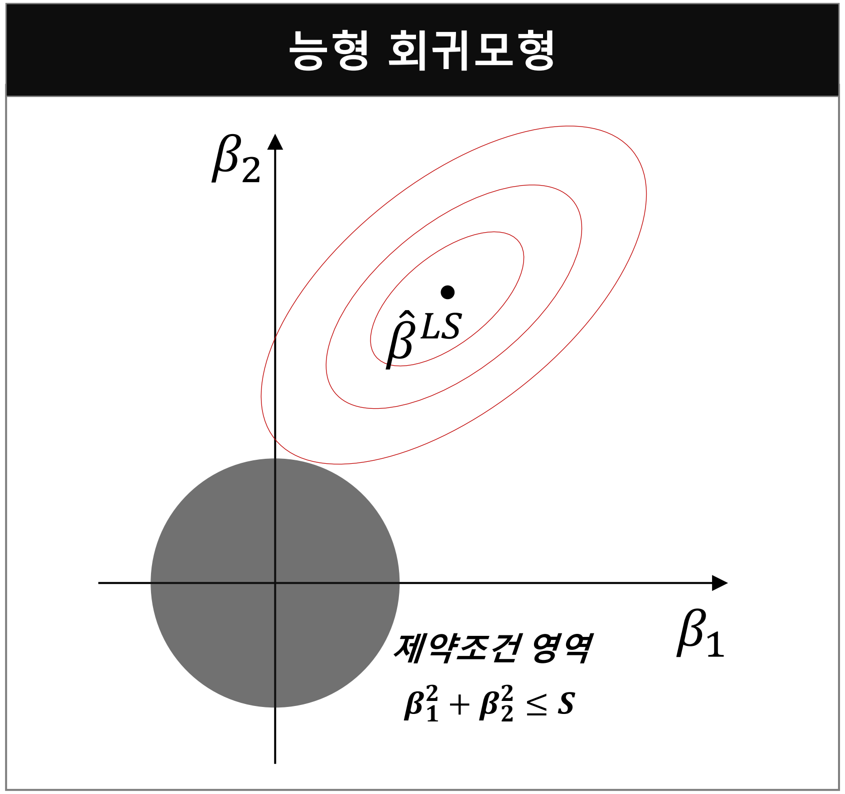
bias

$$\hat{\beta}^{lasso} = \arg\min ||Y - X\beta||^2 - \lambda \sum_{j=1}^{p} |\beta_j|$$

벌점모수(Penalty parameter)
람다(Lambda)를 결정해야 함

Shrinkage Estimator





LASSO(least Absolute Shrinkage and Selection Operator)

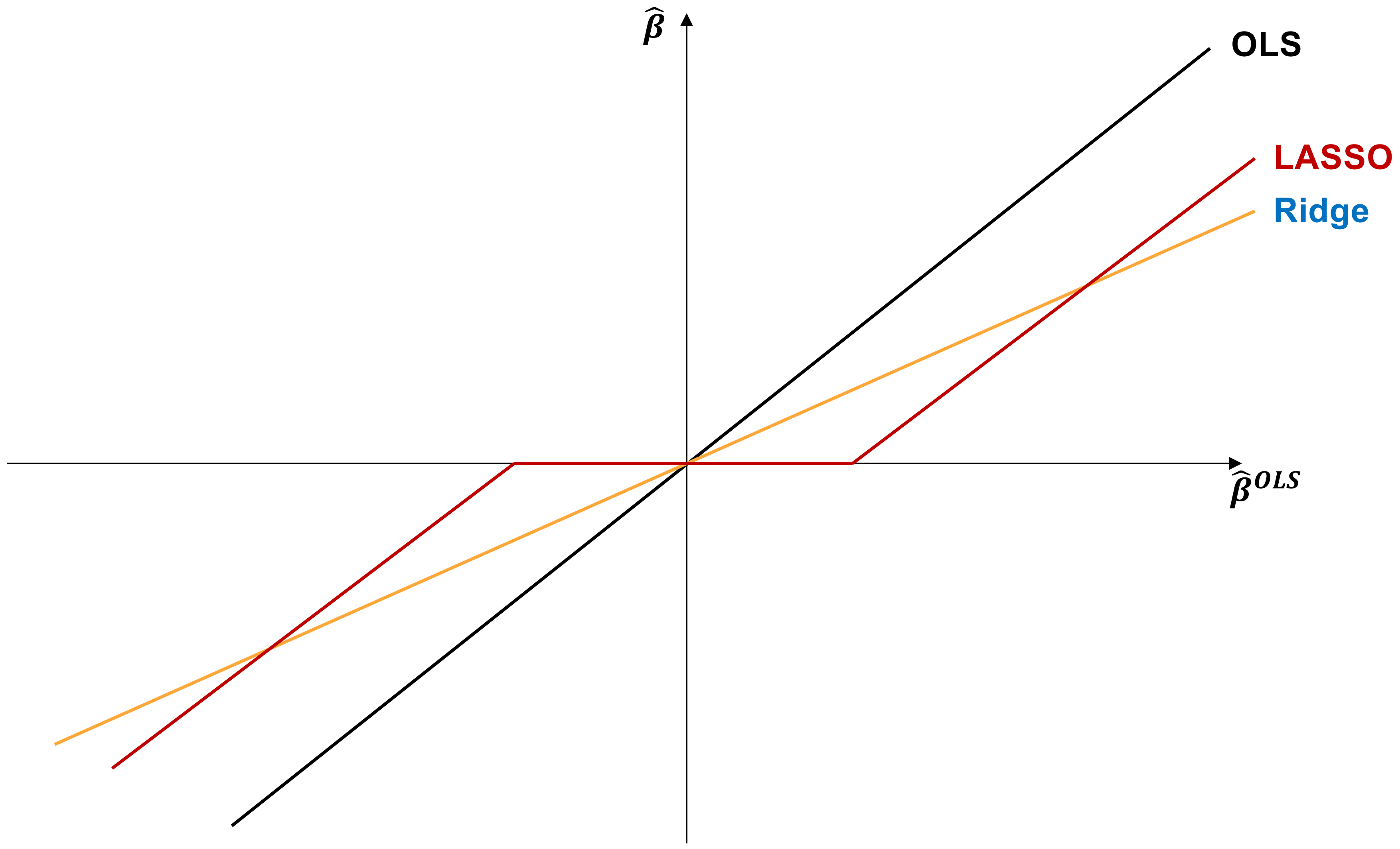
1) 샘플의 개수보다 변수의 개수가 많은 (n < p) 고차원 데이터 분석에 적합함
2) Sparse solution을 제공해줌(유의미하지 않은 회귀계수의 값을 exactly zero로 만들어 줌)
3) 2)에 의해 변수선택의 기능이 있음(Ridge와 다른 점임)
4) 추정된 회귀계수의 값이 OLS의 회귀계수에 비해 0에 가까운 값을 지님(Shrinkage estimator)
5) Logistic Model에서도 적용 가능함
6) 모형에 대한 특정 가정이 존재하지 않음(가령, 오차의 정규성 등)

K-fold cross-validation

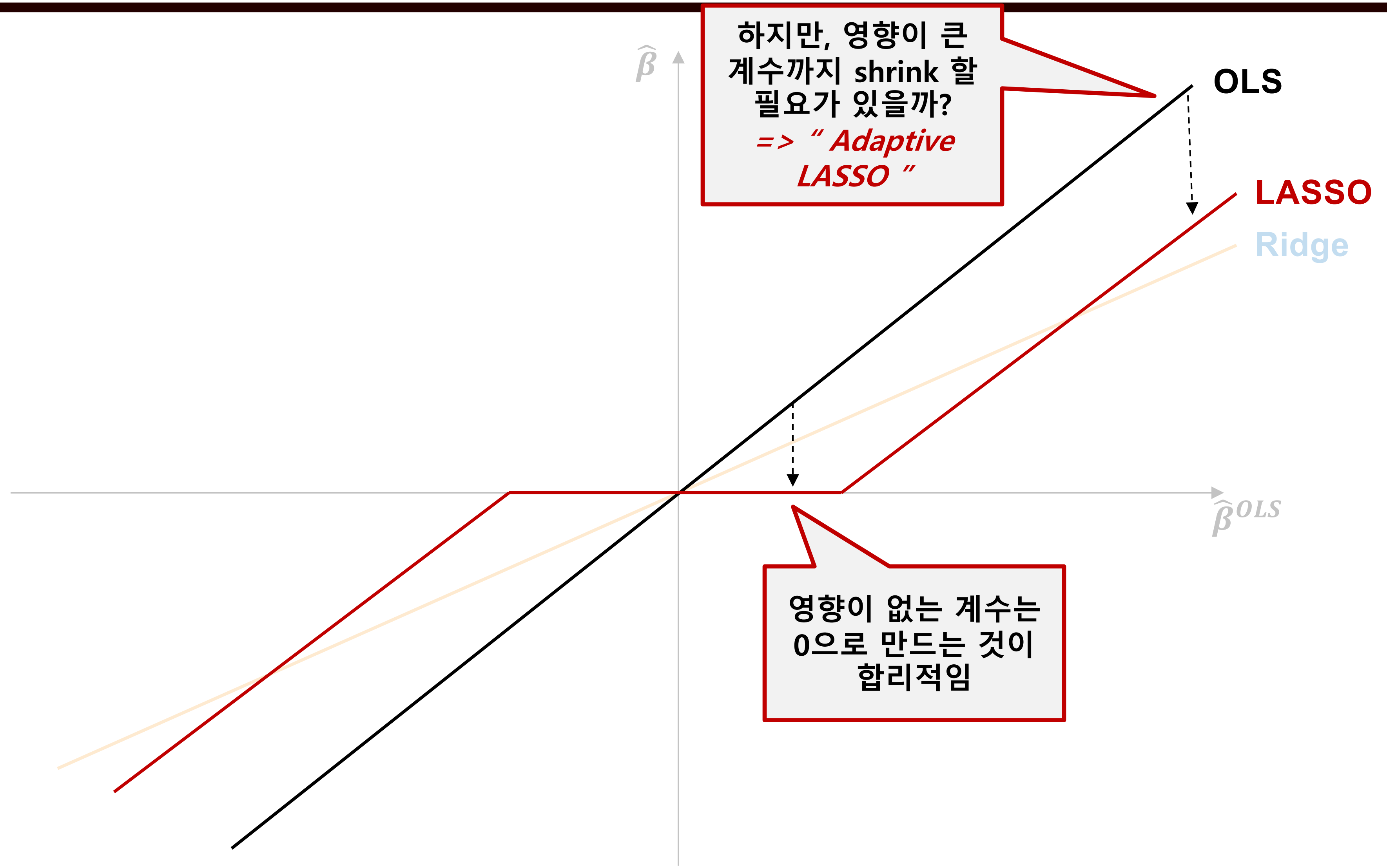
1) 데이터를 random하게 k개의 같은 크기로 쪼갬
2) K-1-fold의 데이터를 트레이닝 데이터로 이용하여 모형을 학습시킴
3) 나머지 데이터(hold-out set)의 Y값을 예측(Prediction)
4) 1~3번 과정을 K번 반복해 모든 Y값에 대한 예측값(Predicted value)을 찾아냄
5) 1~4번까지 과정을 각각의 후보모형마다 실행함
6) 1-5까지 과정을 반복함
7) MSE를 계산하고, 이를 최소화하는 모형 혹은 벌점모수를 선택함

예) 5-fold cross-validation

회귀계수 값들의 비교



회귀계수 값들의 비교



$\widehat{\beta}$
하지만, 영향이 큰 계수까지 shrink 할 필요가 있을까?
=> " Adaptive LASSO "
OLS
LASSO
Ridge
$\widehat{\beta}^{OLS}$
영향이 없는 계수는 0으로 만드는 것이 합리적임

LASSO의 한계와 극복방안

Case#1 -> Adaptive LASSO

✓ OLS는 Consistent estimator인데 반해, Lasso는 아님

✓ 즉, 샘플 수가 무한히 늘어나도 유의미한 변수를 선택하지 못하거나 무의미한 변수를 선택할 수 있음

-> **유의미한 변수의 경우, shrinkage 정도를 줄여줘야 함!**

✓ **Adaptive LASSO**

$$\hat{\boldsymbol{\beta}}^{alasso} = \arg\min ||\boldsymbol{Y} - \boldsymbol{X\beta}||^2 - \boldsymbol{\lambda} \sum_{j=1}^{p} \hat{\boldsymbol{w}}_j |\boldsymbol{\beta}_j| \quad , \quad \hat{\boldsymbol{w}}_j = \frac{1}{\hat{\beta}^{ols}}$$

OLS에서 계수값이 클수록, Weight는 작아져 penalty를 적게 받음

Case#2 -> Elastic Net

✓ 선택할 수 있는 최대 변수의 수 = Sample의 수

✓ 변수가 그룹화 되어 있을 경우, LASSO는 무시함

✓ 변수 간 Correlation 높은 경우, LASSO의 Solution path는 매우 불안정함