Linear Model Selection and Regularization

An Introduction to Statistical Learning

황성원

선형과 비선형 모델의 차이 → 선형 확장!

표준 선형 모델

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

선형 모델

해석 측면에서 유리

선형 모델의 확장!

1. 예측 정확도 향상
2. 모델 해석 용이

비 선형 모델

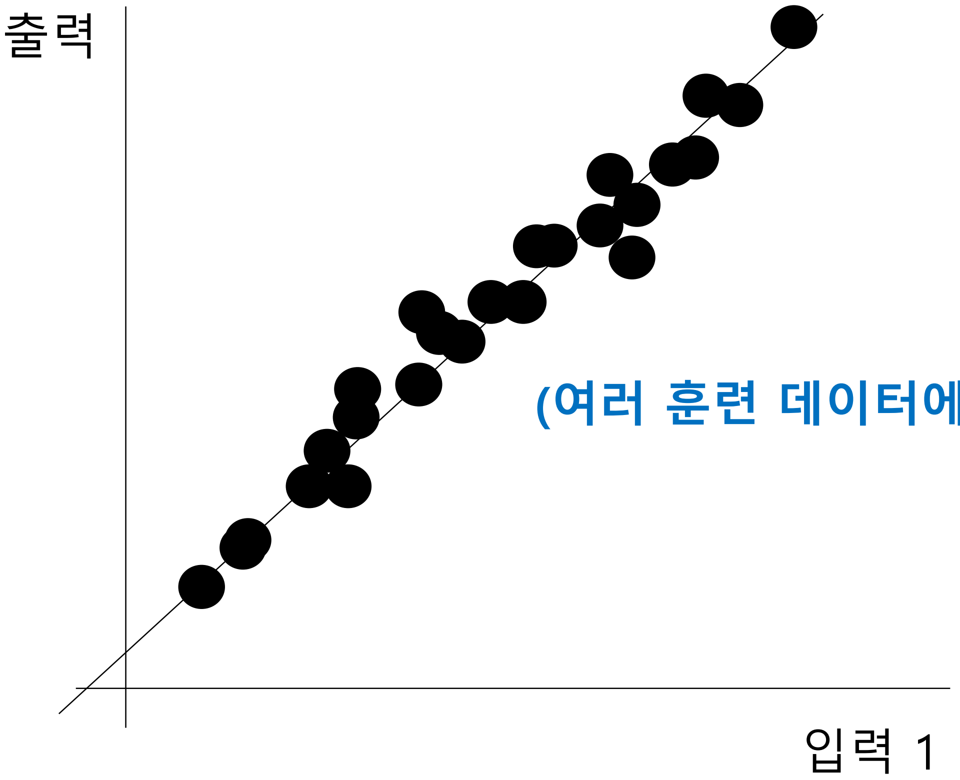
복잡성에서 유리

이 후 장에서 다룸

선형모델 확장 이유1: 예측 정확도 향상

선형 모델의 확장!

1. 예측 정확도 향상 2. 모델 해석 용이



샘플 수 입력 종류 수

$n >> p$, 실제 입출력 관계: 선형

낮은 분산

(여러 훈련 데이터에서 모델이 얼마나 변하는지 정도)

Test 관찰 값에서 높은 예측력!

낮은 바이어스

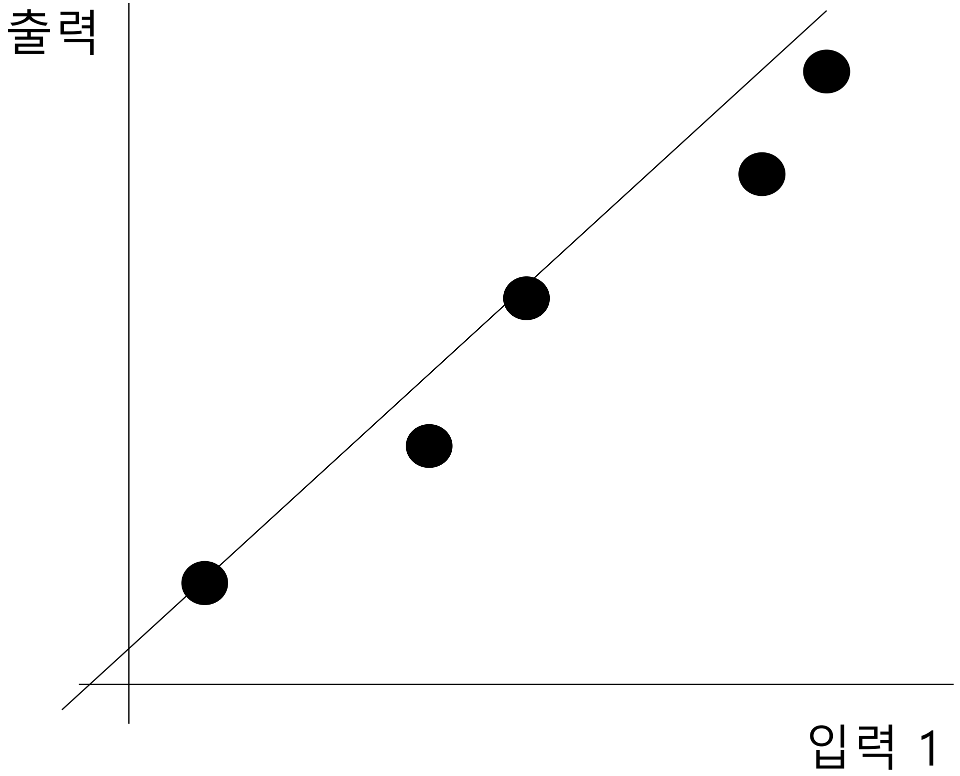
(모델이 얼마나 실제 값을 잘 맞추냐 정도)

SW

선형모델 확장 이유1: 예측 정확도 향상

선형 모델의 확장!

1. 예측 정확도 향상 2. 모델 해석 용이



n > p, 실제 입출력 관계: 선형

↓ 높은 분산 + Overfitting

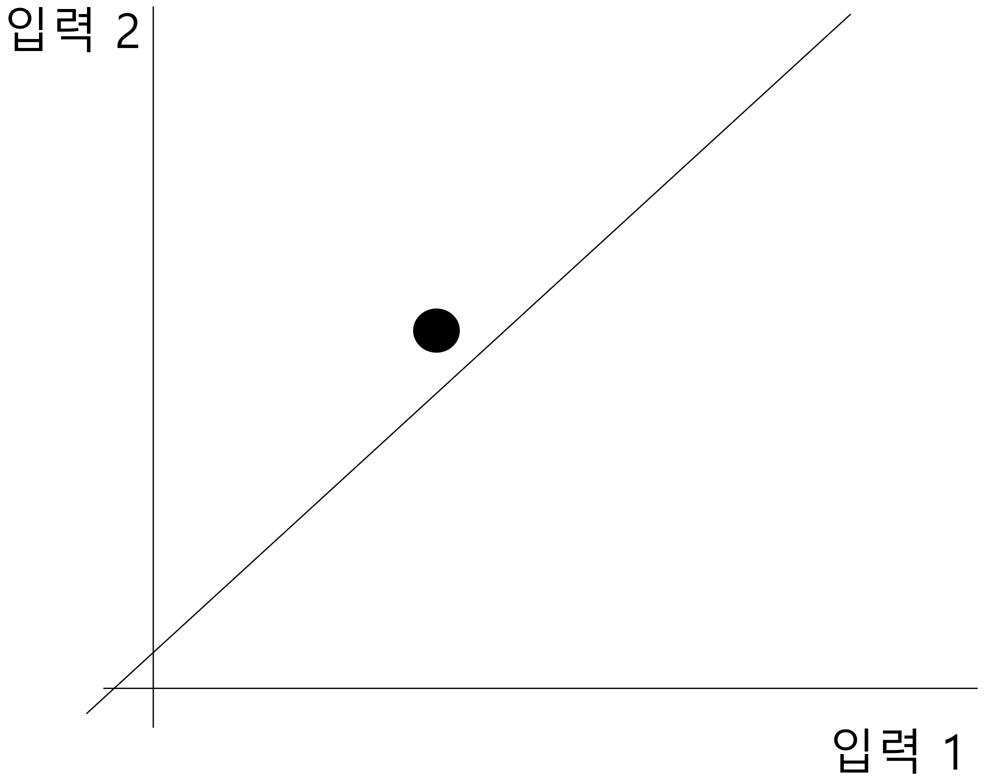
↓ Test 관찰 값에서 낮은 예측력!

↓ 낮은 바이어스

선형모델 확장 이유1: 예측 정확도 향상

선형 모델의 확장!

1. 예측 정확도 향상 2. 모델 해석 용이



$n < p$, 실제 입출력 관계: 선형

분산이 무한대가 된다

유일한 최소자승 계수 못 구함

Test 관찰 값에서 예측 불가

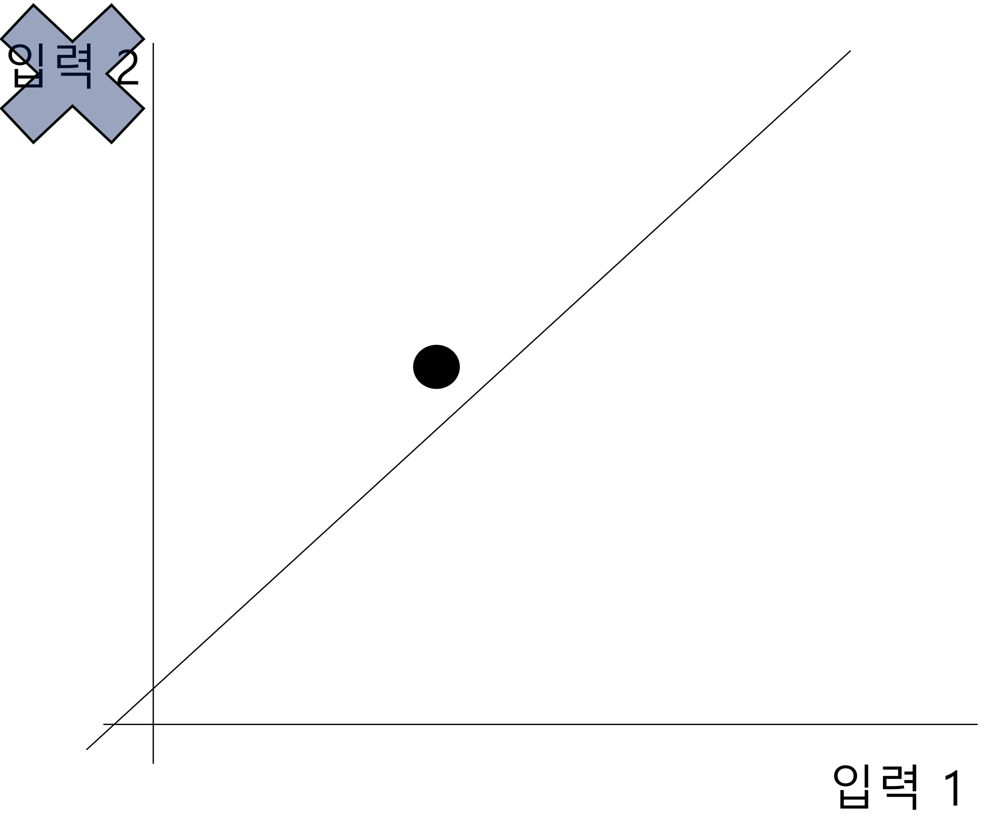
낮은 바이어스

SW

선형모델 확장 이유1: 예측 정확도 향상

선형 모델의 확장!

1. 예측 정확도 향상 2. 모델 해석 용이



$n < p$, **실제 입출력 관계: 선형**

분산이 무한대가 된다

~~유일한 최소자승 계수 못 구함~~

Test 관찰 값에서 예측 가능

낮은 바이어스

선형모델 확장 이유2: 모델 해석 용이

선형 모델의 확장!

1. 예측 정확도 향상
2. **모델 해석 용이**

표준 선형 모델

$$Y = \beta_0 + \beta_1 X_1 + \cancel{\cdots + \beta_p X_p} + \epsilon$$

입출력 사이의 관계가 단순해져 서로 간의 관계를 더욱 쉽게 파악!

→ 모델 해석 용이!

p(입력 종류 수) 줄이는 방법 (1/3)

1. 부분집합 선택(Subset Selection)

표준 선형 모델

$$Y = \beta_0 + \beta_1 \cancel{X_1} + \cdots + \beta_p X_p + \epsilon$$

출력과 관계 없는 입력 변수X를 없애는 방법으로

최종적으로 줄어든 입력들로 최소자승 Fitting을 수행!

SW

p(입력 종류 수) 줄이는 방법 (2/3)

2. Shrinkage or Regularization

표준 선형 모델

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \cancel{\beta_p} X_p + \epsilon$$

1. 부분집합 선택
   (Subset Selection)

계수를 줄이거나, 0으로 정확히(최소자승법으론 불가능) 수렴 시킨다

모든 p개의 입력들로 Fitting을 수행하지만, 최소자승이 아닌 다른 형태를 사용!

p(입력 종류 수) 줄이는 방법 (3/3)

3. 차원 축소 (Dimension Reduction)

표준 선형 모델

p차원 입력 공간

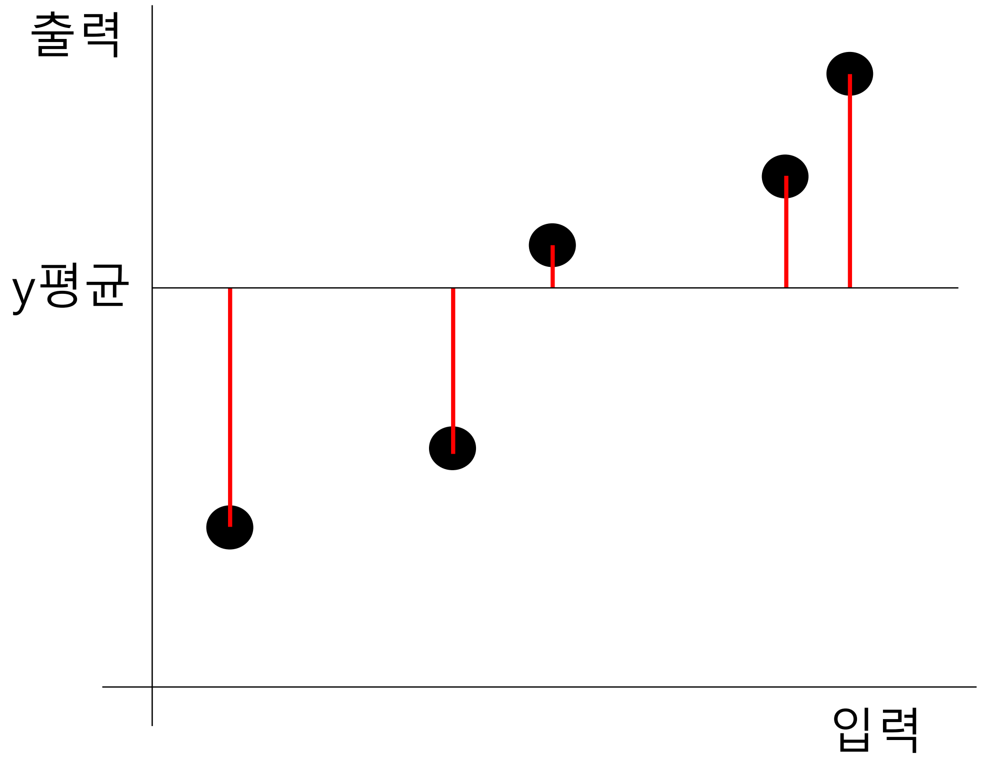
$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

M차원 입력 공간

$$Y = \beta_0 + \beta_1 V_1 + \ldots + \beta_M V_M + \varepsilon$$

복습 – TSS (Total Sum of Squares)

$$\text{TSS} = \sum (y_i - \bar{y})^2$$

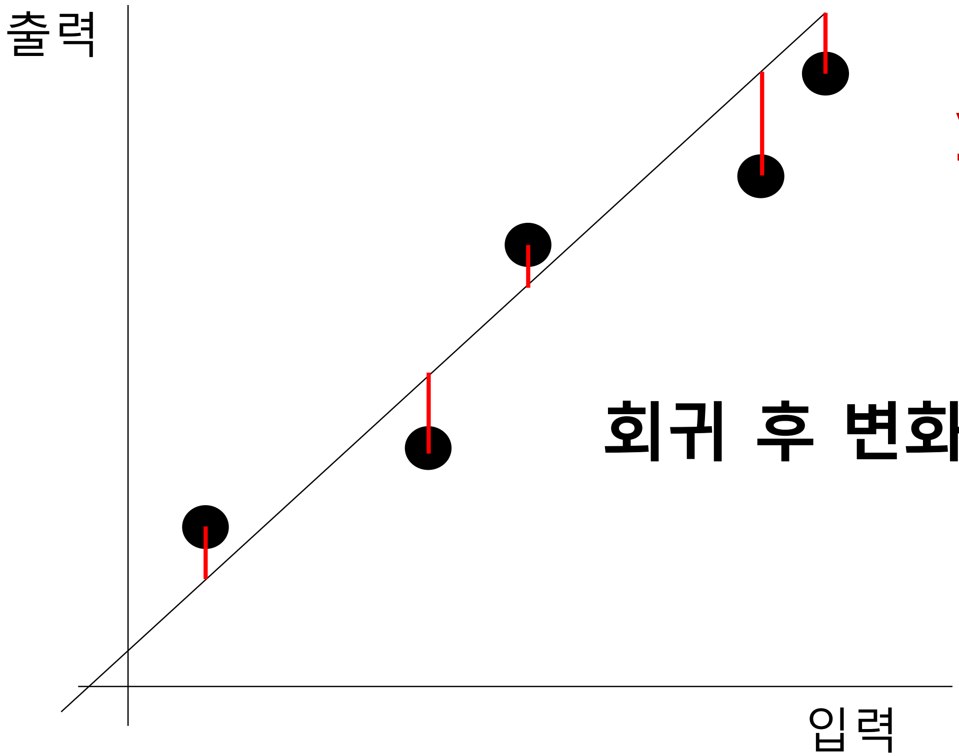


회귀 전 출력 y 자체에 내재된 변화가능성 or 출력 y의 분산

복습 – RSS (Residual Sum of Squares)

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

$$\mathrm{RSS} = (y_1 - \hat\beta_0 - \hat\beta_1 x_1)^2 + (y_2 - \hat\beta_0 - \hat\beta_1 x_2)^2 + \ldots + (y_n - \hat\beta_0 - \hat\beta_1 x_n)^2$$



y출력의 단위로 값이 정해지므로, 기준이 애매

회귀 후 변화가능성 → 회귀식의 정확도를 판별할 때 사용

RSS ↓ 회귀식 정확도 ↑

복습 – R^2 결정계수

RSS에서의 판별 문제점: y출력의 단위로 값이 정해지므로, 기준이 애매

⬇

TSS와 RSS의 상대적 비를 통해 0과 1사이로 스케일링을 옮겨서 그 정도를 파악!

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

RSS ⬇ R^2 ⬆ 회귀식 정확도 ⬆

1. 부분집합 선택(Subset Selection)

1. **Best Subset Selection**

2. **Forward Stepwise Selection**

3. **Backward Stepwise Selection**

1. Best Subset Selection:

모든 부분집합들 중 가장 좋은 것으로 하겠다

p=1인 경우,

$$Y = \beta_0 + \beta_1 X_1$$

$$Y = \beta_0 + \beta_2 X_2$$

$$\cdots$$

$$Y = \beta_0 + \beta_p X_p$$

$\binom{p}{1} = {}_pC_1$ **= 총 p개의 부분집합이 생김!**

SW

1. Best Subset Selection:

모든 부분집합들 중 가장 좋은 것으로 하겠다

p=2인 경우,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 \qquad {}_pC_2$$ **= 총 p(p-1)/2 개의 부분집합**

$$\cdots$$

$$Y = \beta_0 + \beta_{p-1} X_{p-1} + \beta_p X_p$$

1. Best Subset Selection:

모든 부분집합들 중 가장 좋은 것으로 하겠다

p=k인 경우, $_{p}C_{k}$

p=1부터 p까지 모든 경우의 수: p가 클 경우, 계산 량이 엄청나다!

$$X_1 \; X_2 \; X_3 \cdots \; X_p$$

$$2 \times 2 \times 2 \times \ldots \times 2 = 2^p$$

SW

1. Best Subset Selection

Algorithm 6.1 *Best subset selection*

Step 1.

M_0 **: null model(입력이 하나도 없는 경우)**

Sample Mean 값을 예측 한다.

$$Y = \beta_0 = \bar{y}$$

1. Best Subset Selection

Algorithm 6.1 *Best subset selection*

Step 2.

k=1, 2, …, p: 모델은 M_k이라고 표기.

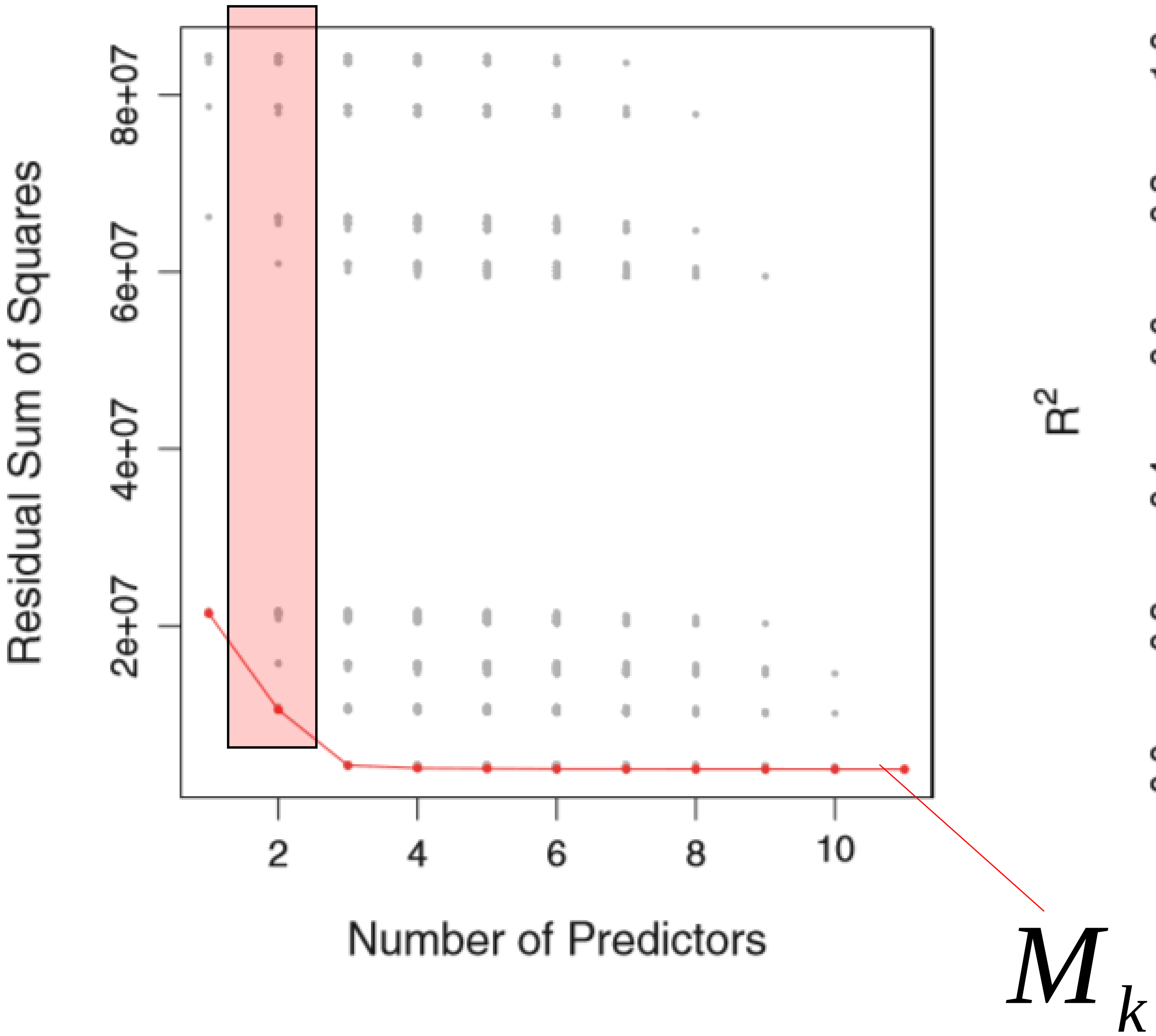
(1) k개의 입력이 포함된 모든 $\binom{p}{k}$개의 Model을 모두 Fitting 한다.

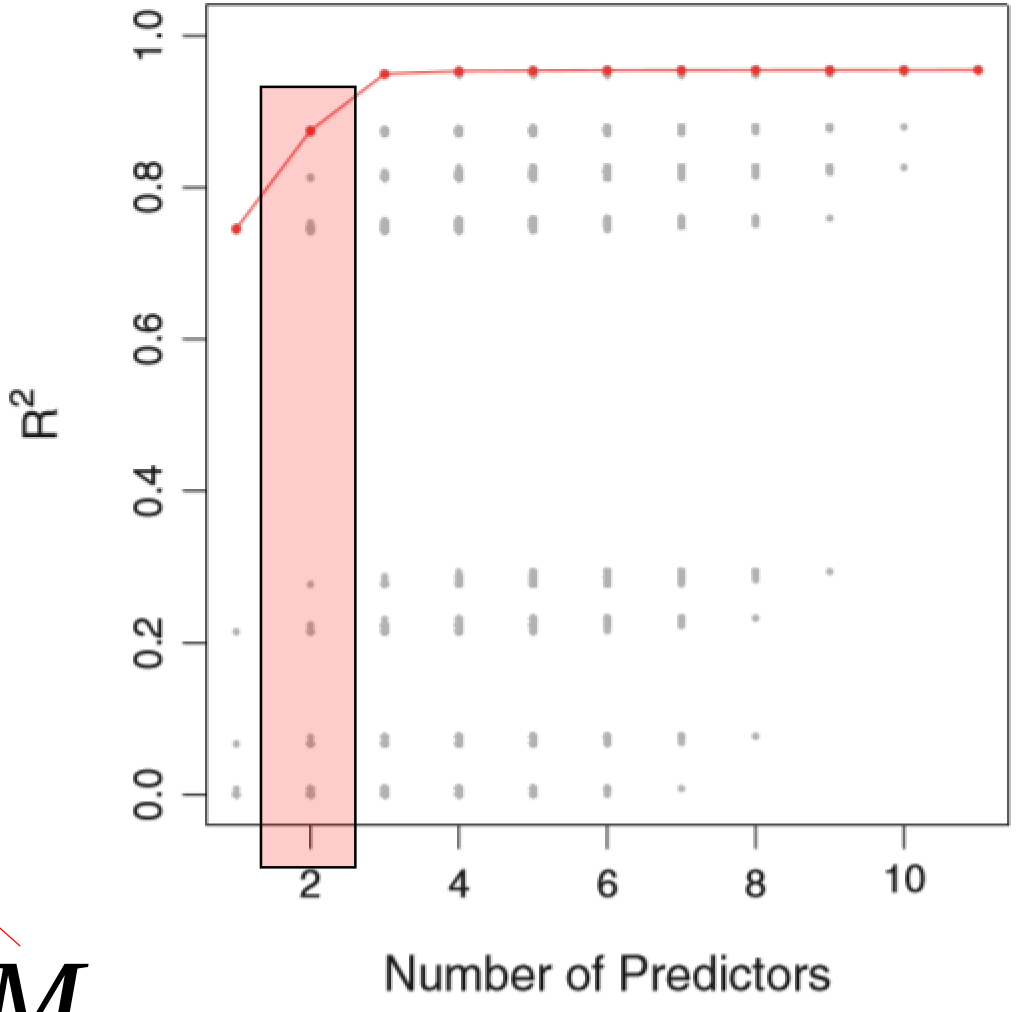
(2) $\binom{p}{k}$개의 Model 중에서 Best를 고른다.

(여기서 Best는 RSS가 가장 작고, R^2는 가장 큰 값이다.

1. Best Subset Selection

Step 2.





SW

Step 2에서의 이슈

훈련 데이터	Test 데이터

RSS ↓ R^2 ↑

훈련 데이터 Error ↓ $\neq$ Test Error ↓

1. 간접적 방법(Adjustment to 훈련 데이터) 2. 직접적 방법(Cross-Validation Approach)

Step 2에서의 이슈 – 간접적 방법

훈련 데이터	Test 데이터

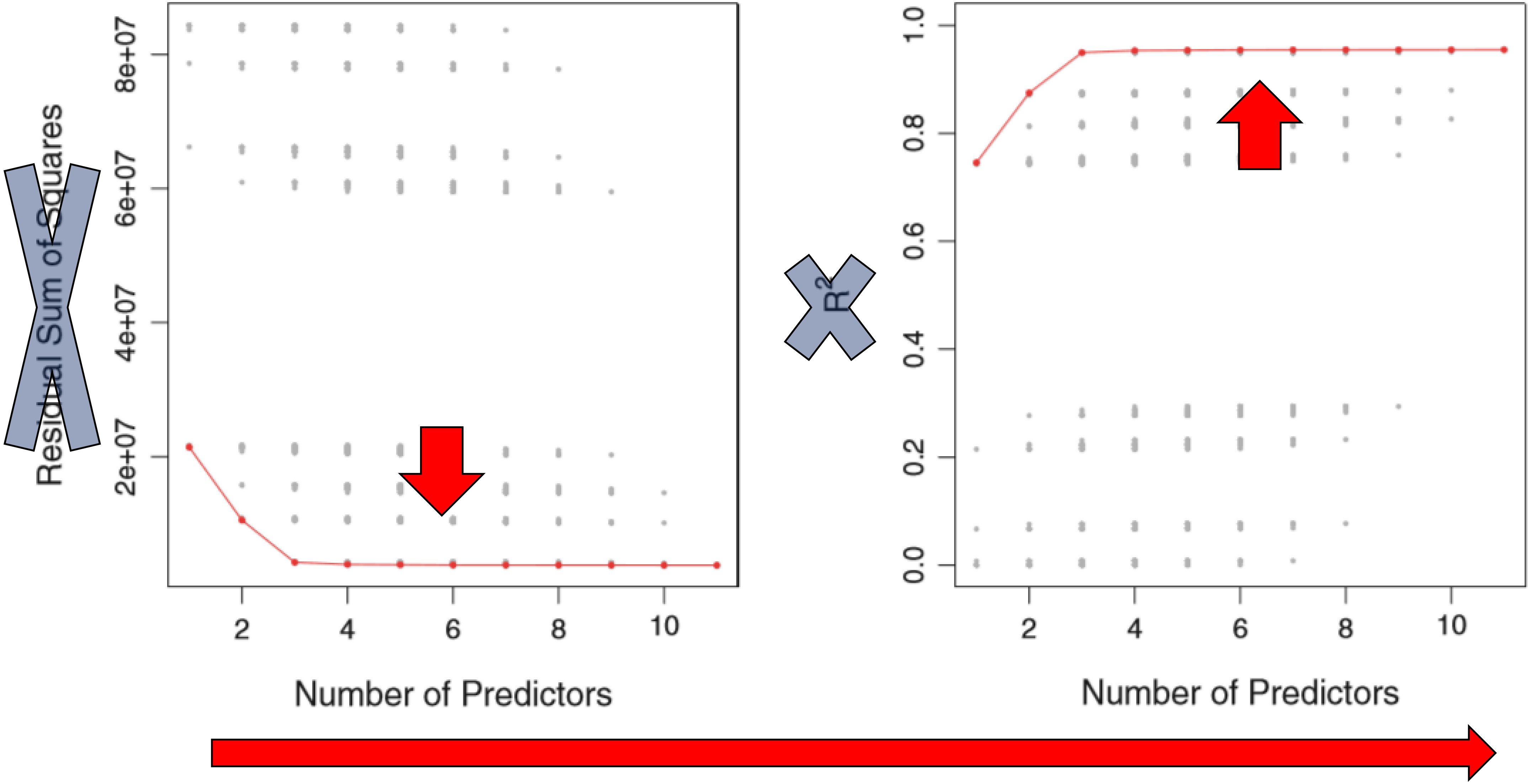
RSS ↓ R^2 ↑

훈련 데이터 Error ↓ $\neq$ Test Error ↓

1. 간접적 방법(Adjustment to 훈련 데이터) 2. 직접적 방법(Cross-Validation Approach)

SW

Training Set 에서는… Test Set에서는 X



사전 지식 For Step 3 – 1. C_p

Training MSE < Test MSE

$$C_p = \boxed{\frac{1}{n}\left(\mathrm{RSS}\right.}\boxed{\left.+\,2d\hat{\sigma}^2\right)} \approx \textbf{Test MSE의 추정치}$$

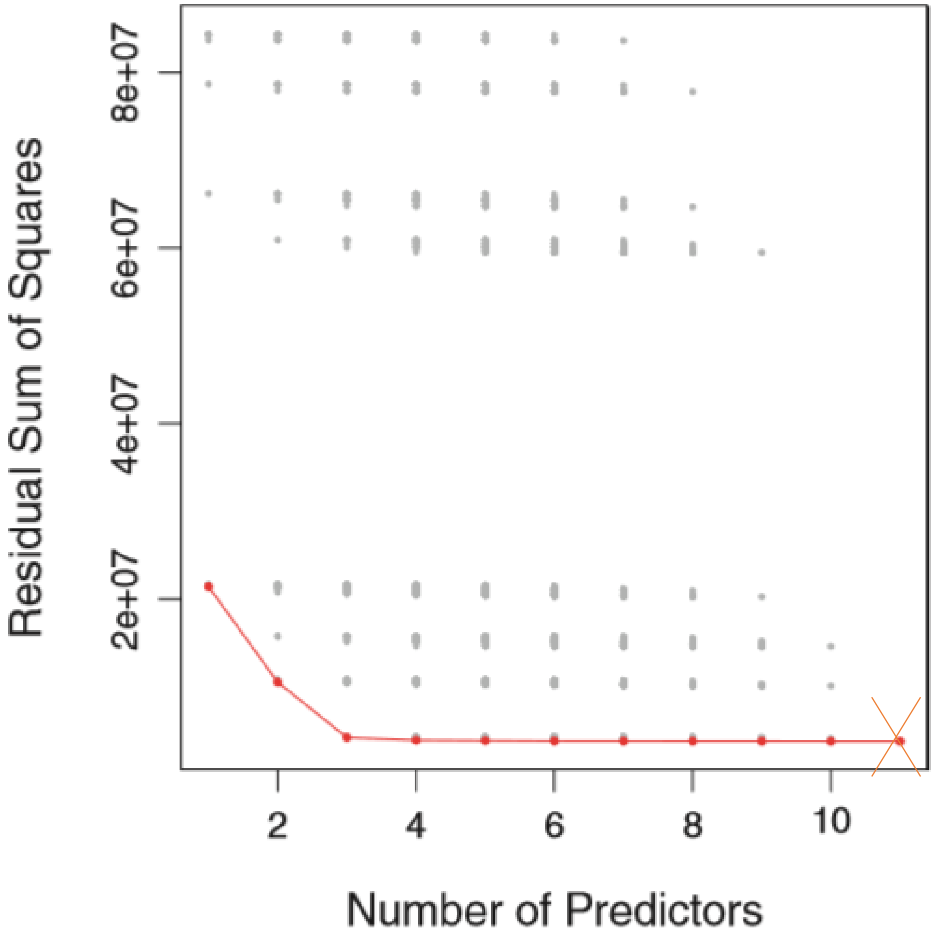
Penalty

d: 모델이 포함하고 있는 입력 종류의 수

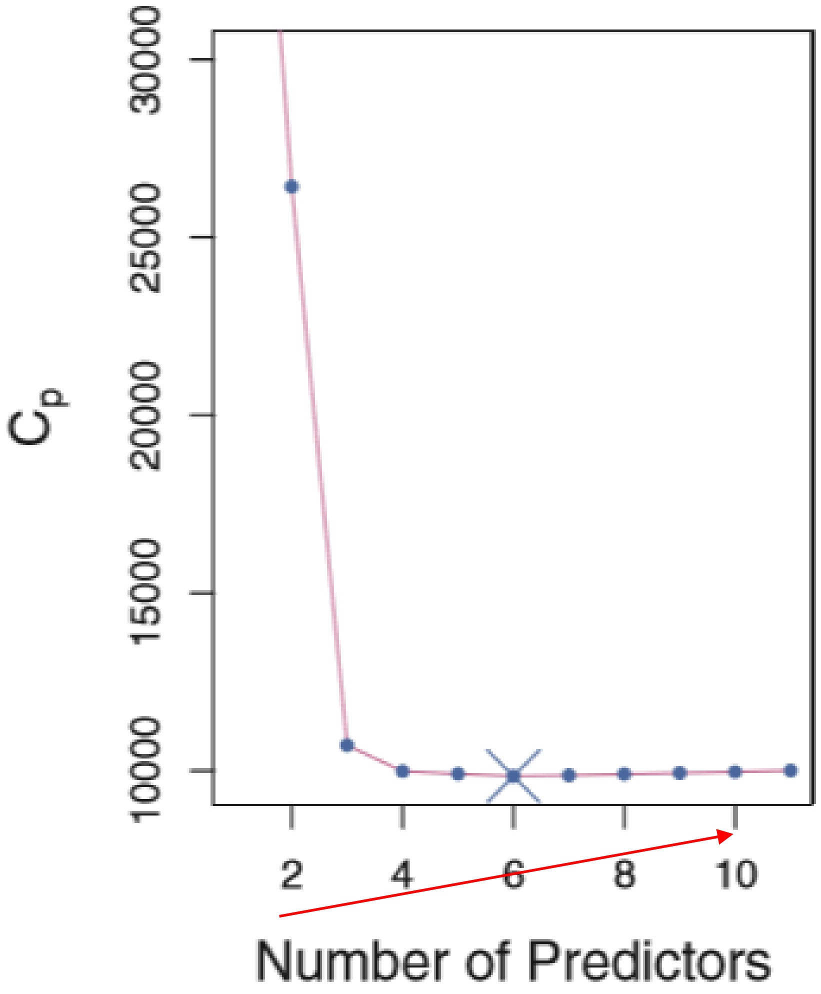
$\hat{\sigma}^2$ **: Irreducible Error의 분산의 추정치**

SW

사전 지식 For Step 3 – 1. $C_p = \frac{1}{n}\left(\text{RSS} + 2d\hat{\sigma}^2\right)$



VS.



SW

사전 지식 For Step 3 – 2. AIC (Akaike Information Criterion)

Maximum Likelihood → Model Fitting

Irreducible Error = Gaussian Error → ML = Least Squres

$$\mathrm{AIC} = \frac{1}{n\boxed{\hat{\sigma}^2}}\left(\mathrm{RSS} + 2d\hat{\sigma}^2\right)$$

사전 지식 For Step 3 – 3. BIC (Bayesian Information Criterion)

Least Squares Model + Bayesian Point of View

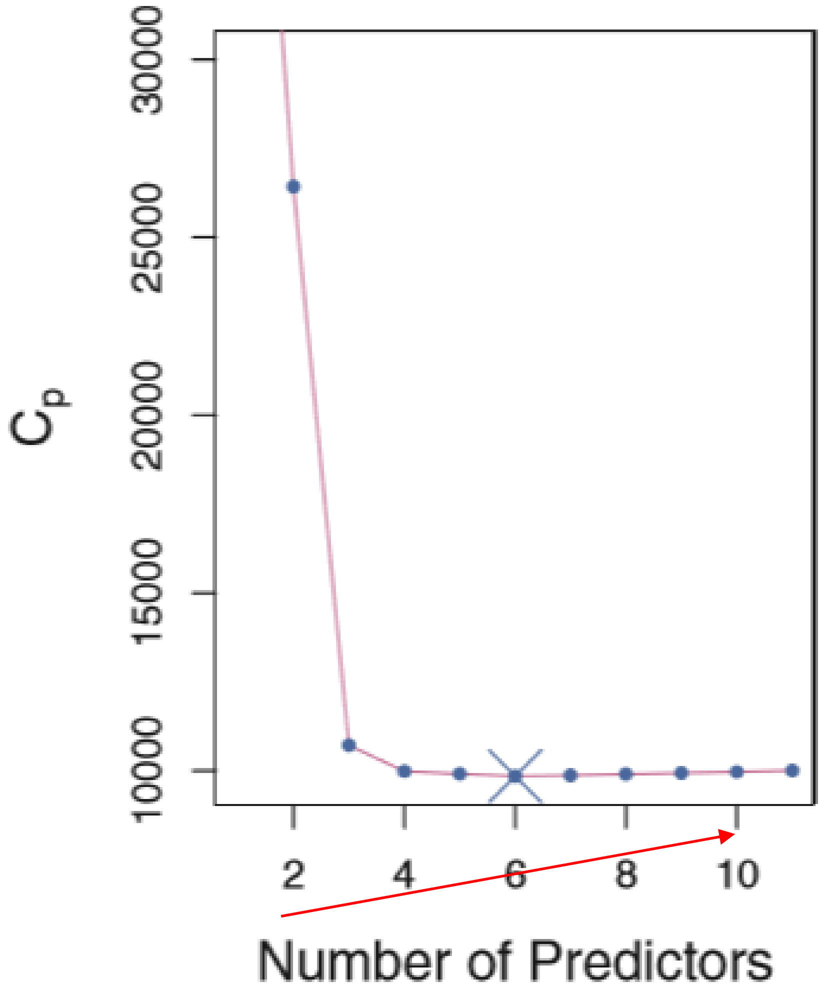
⬇

$$\mathrm{BIC} = \frac{1}{n}\left(\mathrm{RSS} + \boxed{\log(n)}\, d\hat{\sigma}^2\right)$$

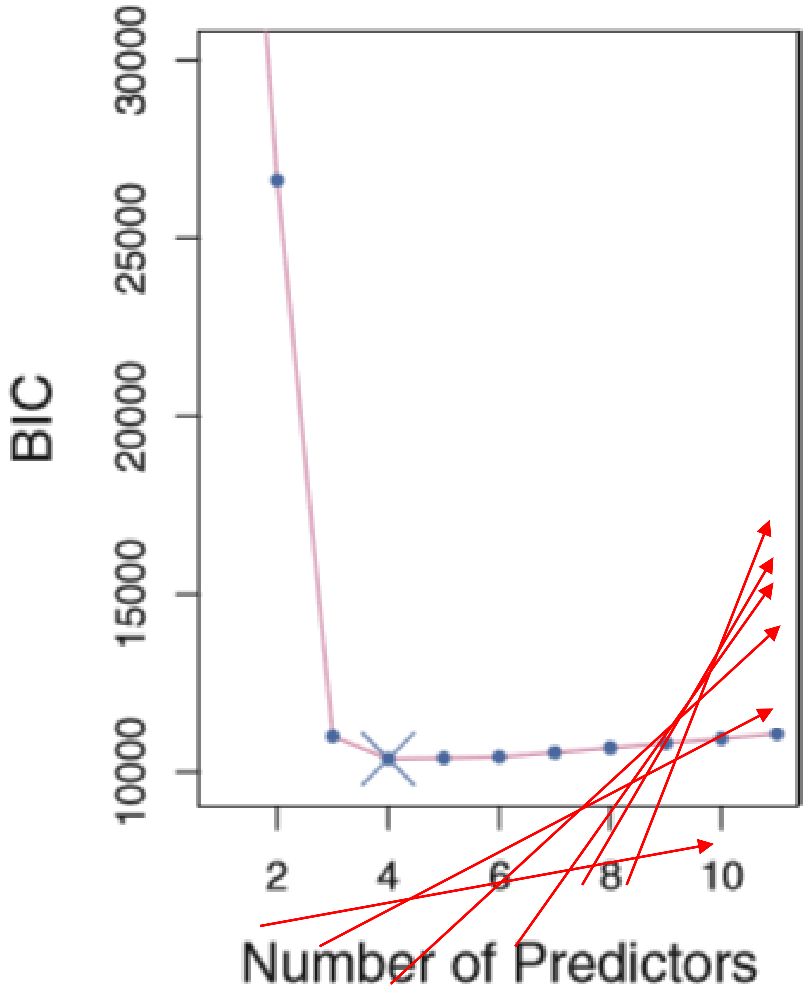
$\log(n) > 2, \quad n > 7$ ➡ Penalty 증가(Cp에 비해)

→ 더 작은 입력 종류에서 최소값

사전 지식 For Step 3 – 3. BIC (Bayesian Information Criterion)



vs.



SW

사전 지식 For Step 3 – 4. Adjusted R^2

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}$$

d 입력 종류를

Fitting 모델에 많이 넣으면 →

$$\frac{\text{RSS}}{n-d-1} \qquad \text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}$$

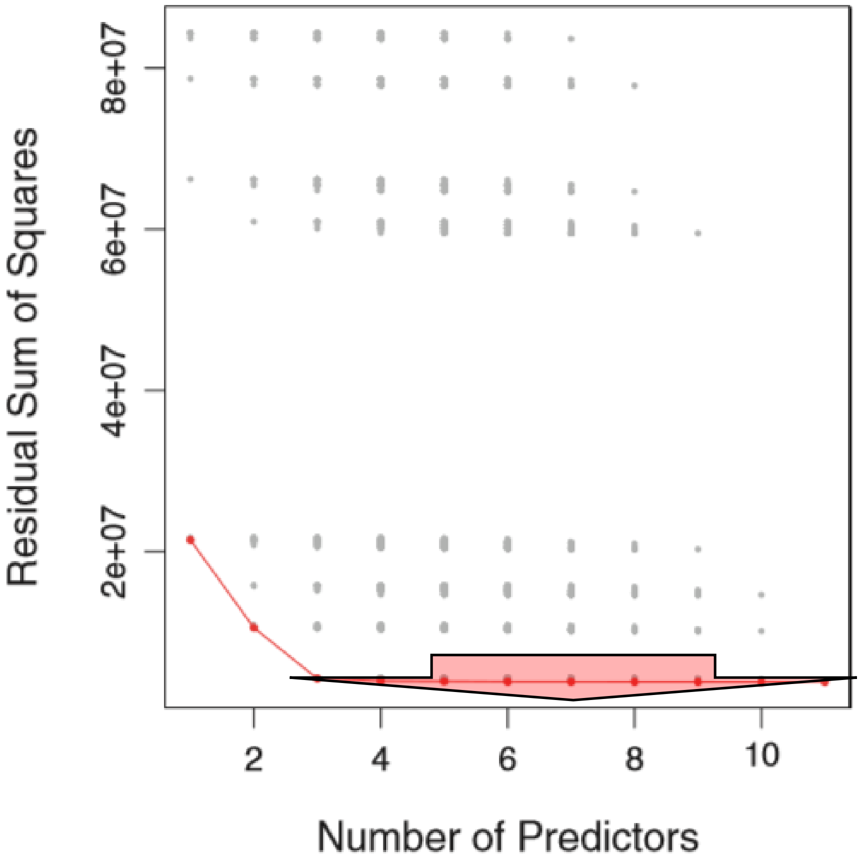
$$\frac{\text{RSS}}{n-d-1} \qquad \text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}$$

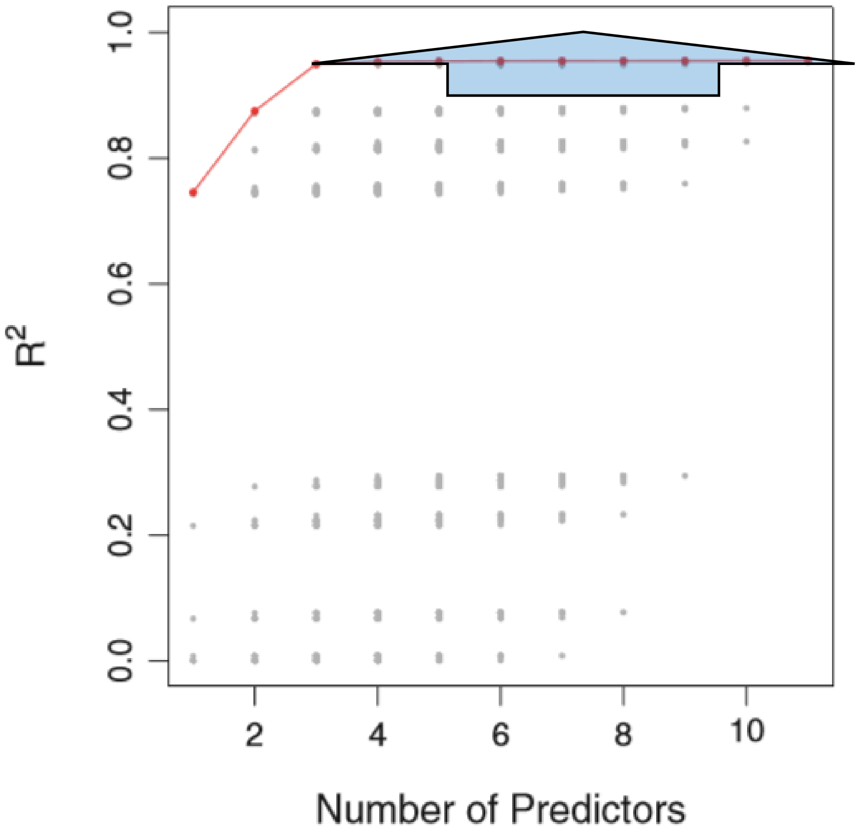
사전 지식 For Step 3 – 4. Adjusted R^2

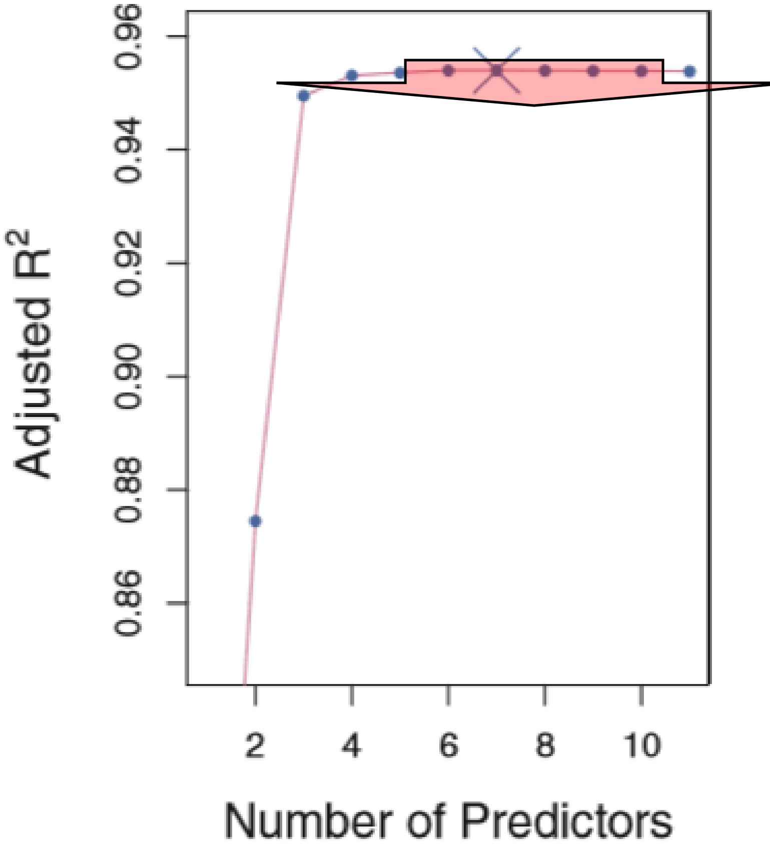
올바른 입력이 모두 들어간 후,

노이즈 입력이 들어갔을 때

$$\frac{\mathrm{RSS}}{n-d-1} = \text{Adjusted } R^2 \quad \text{VS.} \quad R^2$$







사전 지식 For Step 3 – 비교

Adjusted R^2

C_p

AIC

BIC

계산 용이

사전 지식 For Step 3 – 비교

Adjusted R^2

C_p

AIC

BIC

Cross-Validation Approach

계산 용이

가정이 적고, 직접 Error 추정

사전 지식 For Step

컴퓨팅의 발전

과거

Adjusted R^2

C_p

AIC

BIC

계산 용이

현재

Cross-Validation Approach

가정이 적고, 직접 Error 추정

SW

Step 2에서의 이슈 – 직접적 방법

훈련 데이터 | Test 데이터

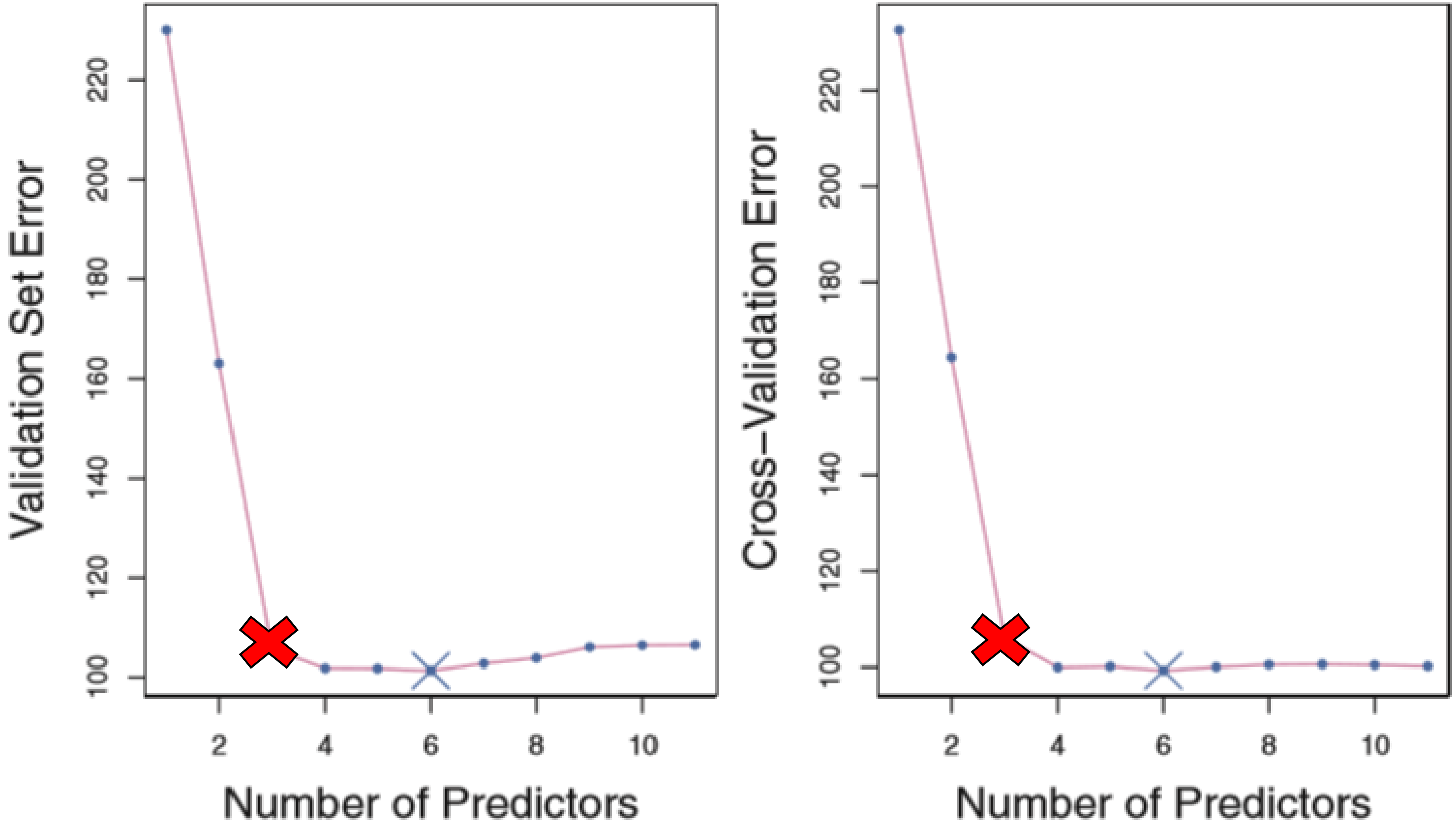
RSS R^2

훈련 데이터 Error $\neq$ Test Error

1. 간접적 방법(Adjustment to 훈련 데이터) **2. 직접적 방법(Cross-Validation Approach)**

직접적 방법 – Cross-Validation Approach

One Standard Error Rule



SW

1. Best Subset Selection

Algorithm 6.1 *Best subset selection*

Step 3.

$M_0, \ldots, M_p$ **중에서 하나의 Best Model을 구한다.**

아래의 기준들을 이용해서…

Adjusted R^2 C_p **AIC BIC Cross-Validation Approach**

SW

1. Best Subset Selection 문제점

1. 컴퓨팅 부담이 너무 큼

2. 통계적으로 p가 너무 클 경우,

Fitting 해야 할 공간이 넓어 지고, 이는 Training Data에만 좋아 보이는 지점

에 Fitting 하여 Test 공간에서는 좋지 못한 결과를 주는 Overfitting의

가능성이 높아지고, 계수의 추정치의 분산이 높아 짐

2. Forward Stepwise Selection

Algorithm 6.2 *Forward stepwise selection*

Step 1.

M_0 **: null model(입력이 하나도 없는 경우)**

2. Forward Stepwise Selection

Algorithm 6.2 *Forward stepwise selection*

Step 2.

k = 0, … , p-1:

(1) M_k **안에 하나의 추가적인 입력을 넣은 p-k 개의 모델을 고려해라.**

(2) p-k개의 모델에서 가장 작은 RSS와 가장 큰 R^2 **값을 보이는 모델을**

Best로 지정하고 이를 M_{k+1}**라 칭한다.**

2. Forward Stepwise Selection

Algorithm 6.2 *Forward stepwise selection*

Step 3.

$M_0, \ldots, M_p$ 중에서 최종 **Best Model**을 구한다. (아래 기준 이용)

Adjusted R^2	C_p	**AIC**	**BIC**	**Cross-Validation Approach**

2. Forward Stepwise Selection

컴퓨팅 이점은 얼마나? – 총 경우의 수

Forward Stepwise Selection

$$1+\sum_{k=0}^{p-1}(p-k)=1+\frac{p(p+1)}{2}$$

Best Subset Selection

$$2^p$$

$$p=20,$$

Forward Stepwise Selection

211

Best Subset Selection

1,048,576

SW

2. Forward Stepwise Selection – 헛점

Best Subset Selection

$$p = 1, \qquad X_1$$

$$p = 2, \qquad X_2 \; X_3$$

Forward Stepwise Selection

$$p = 1, \qquad \boxed{X_1}$$

$$p = 2, \qquad \boxed{X_1} \; X_?$$

Best Model 추정 실패!

SW

3. Backward Stepwise Selection

Algorithm 6.3 *Backward stepwise selection*

Step 1.

M_p **: full model(모든 p입력이 포함된 모델)**

3. Backward Stepwise Selection

Algorithm 6.3 *Backward stepwise selection*

Step 2.

k = p, p-1, … , 1:

(1) M_k **안에 하나의 추가적인 입력을 뺀 k-1개의 입력을 가진**

k 개의 모델을 고려해라.

(2) k개의 모델에서 가장 작은 RSS와 가장 큰 R^2 **값을 보이는 모델을**

Best로 지정하고 이를 M_{k-1} **라 칭한다.**

SW

3. Backward Stepwise Selection

Algorithm 6.3 *Backward stepwise selection*

Step 3.

$M_0, \ldots, M_p$ **중에서 최종 Best Model을 구한다. (아래 기준 이용)**

Adjusted R^2 C_p **AIC BIC** **Cross-Validation Approach**

3. Backward Stepwise Selection – 헛점

Best Subset Selection

$$p = 1, \qquad X_1$$

$$p = 2, \qquad X_2 \;\; X_3$$

Backward Stepwise Selection

$$p = 2, \qquad \boxed{X_2} \;\; X_3$$

$$p = 1, \qquad \boxed{X_2} \qquad ✖$$

Best Model 추정 실패!

Forward VS. Backward

High-Dimensional Setting

$$\boxed{n < p}$$

Forward VS. Backward

$$M_0, \ldots, M_{n-1}$$

Fitting이 시작 될 수 없다

$n \geq p$, **Least Squares**

Unique Solution X

Hybrid Approaches

Forward + Backward

$$\approx$$

Best Subset Selection

p(입력 종류 수) 줄이는 방법 (1/3)

1. 부분집합 선택(Subset Selection)

표준 선형 모델

$$Y = \beta_0 + \beta_1 \cancel{X_1} + \cdots + \beta_p X_p + \epsilon$$

출력과 관계 없는 입력 변수X를 없애는 방법으로

최종적으로 줄어든 입력들로 최소자승 Fitting을 수행!

p(입력 종류 수) 줄이는 방법 (2/3)

2. Shrinkage or Regularization

표준 선형 모델

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \cancel{\beta_p} X_p + \epsilon$$

1. 부분집합 선택
(Subset Selection)

계수를 줄이거나, 0으로 정확히(최소자승법으론 불가능) 수렴 시킨다

모든 p개의 입력들로 Fitting을 수행하지만, 최소자승이 아닌 다른 형태를 사용!

p(입력 종류 수) 줄이는 방법 (3/3)

3. 차원 축소 (Dimension Reduction)

표준 선형 모델

p차원 입력 공간

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

M차원 입력 공간

$$Y = \beta_0 + \beta_1 V_1 + \ldots + \beta_M V_M + \varepsilon$$

Thank you!