

## Chapter06

## 선형회귀

# I기타 회귀분석

M T W T F S S

2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

FASTCAMPUS  
ONLINE

금융공학/퀀트 I

강사. 장순용

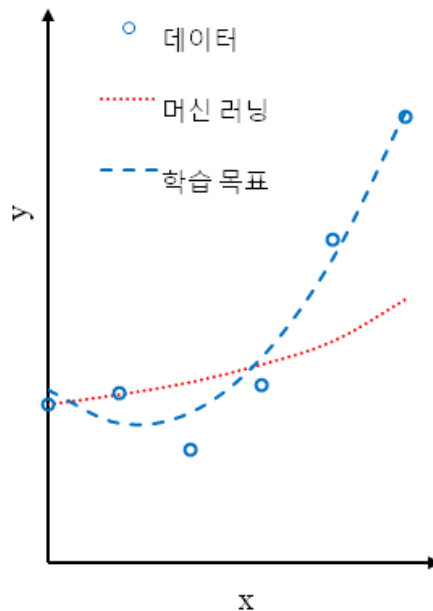
# I 키포인트

- 편향 오류와 분산 오류.
- Ridge 회귀.
- Lasso 회귀.
- 다항식 회귀.
- 푸아송 회귀.



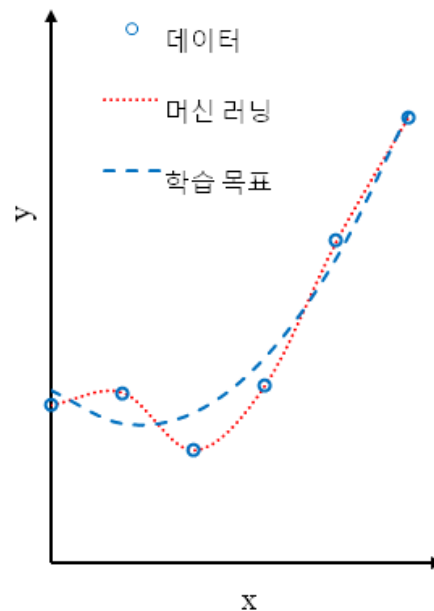
## I 편향 오류 (Bias error)

- 편향 오류 (bias error) 또는 과소적합 오류 (underfitting error).
- 모형이 편향적 즉 과하게 단순해서 발생하는 오류의 유형이다.



## I 분산 오류 (Variance error)

- 분산 오류 (variance error) 또는 과적합 오류 (overfitting error).
- 모형이 과하게 복잡해서 발생하는 오류이며 매개변수 최적화의 어려움으로 표출된다.



## I 분산 오류 (Variance error)

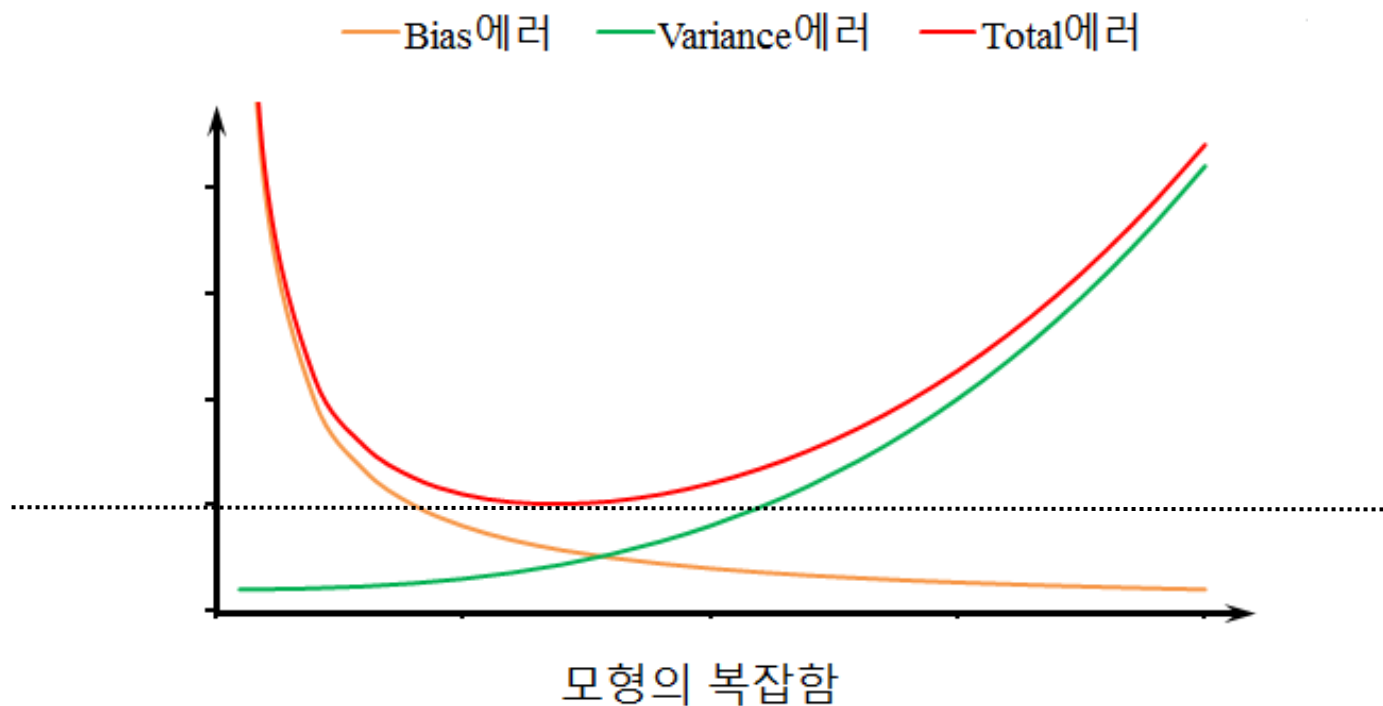
- 분산 오류 (variance error) 또는 과적합 오류 (overfitting error).
- 모형이 과하게 복잡해서 발생하는 오류이며 매개변수 최적화의 어려움으로 표출된다.
- In-sample 오류는 작지만 Out-of-sample 오류는 큰 경우이다.

# I 토탈 오류

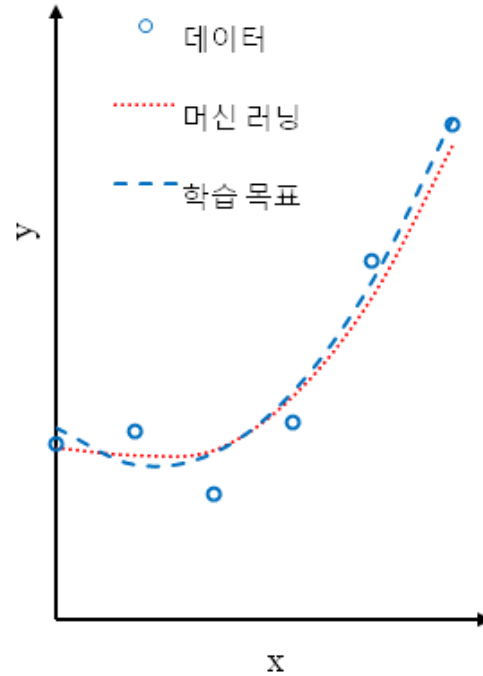
$$\text{토탈 오류} = \text{편향 오류} + \text{분산 오류} + \text{상수}$$

# I Out-of-sample 시험 오류의 최소화

- 모형의 복잡함 (complexity)에는 최적점(optimal point)이 있다.



# I Out-of-sample 시험 오류의 최소화



최적화된 모형



## I Ridge 회귀

- OLS해는  $\|\vec{\epsilon}\|^2$ 를 최소화 하는 계수벡터를 구한다.
- Ridge회귀에서는 다음 손실함수를 최소화 한다. (L2 정규화)

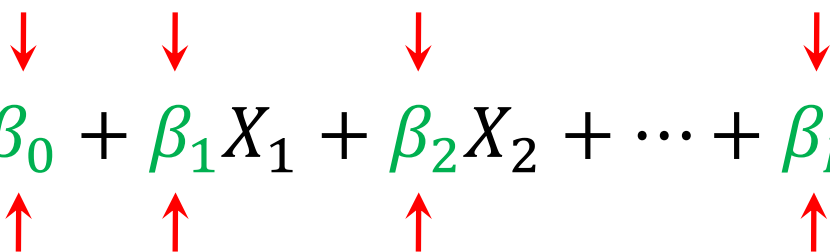
$$L = \|\vec{\epsilon}\|^2 + \lambda \sum_{i=0}^K \beta_i^2$$

- $\lambda$ 는 양수로서 크면 클수록 분산오류를 줄이며 편향오류를 증가시킨다.
- 과적합 (overfitting)의 상황이 의심될 때 사용한다.
- 회귀계수의 절대값은 억제되지만 정확하게 0이 되지는 않는다.

## I Ridge 회귀

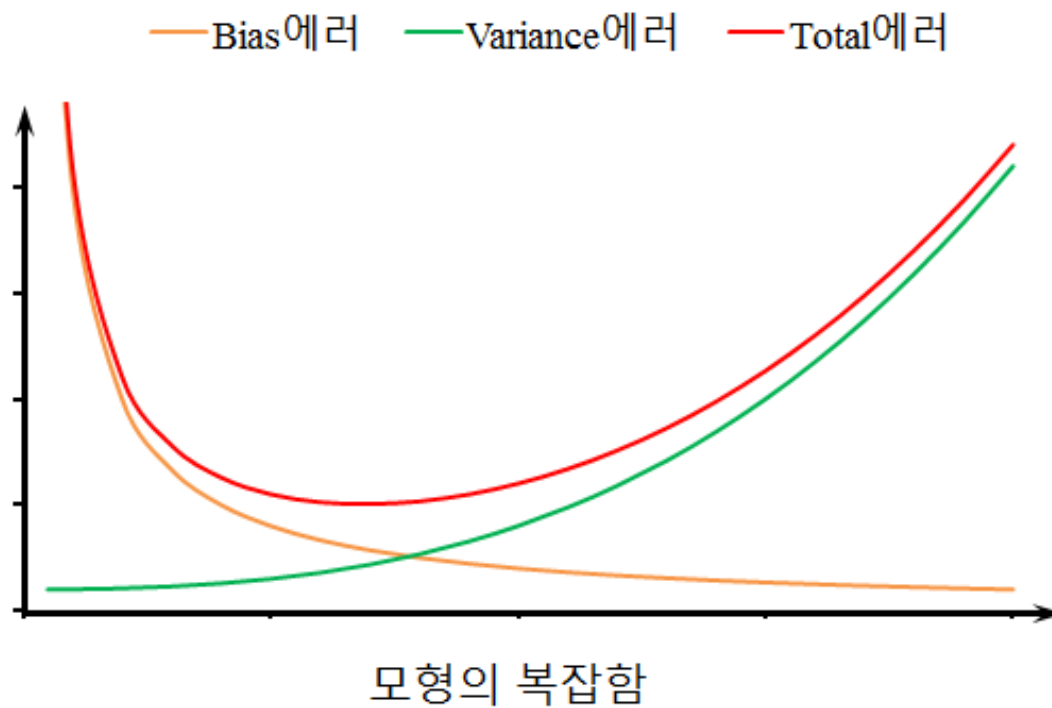
- $\lambda$ 는 크면 클수록 회귀계수의 증가를 억제한다.

$$L = \|\vec{\varepsilon}\|^2 + \lambda \sum_{i=0}^K \beta_i^2$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \varepsilon$$


# I Ridge 회귀

- 편향오류와 분산오류 사이의 trade-off 관계를 상기해 본다.



# I Lasso 회귀

- Lasso회귀에서는 다음 손실함수를 최소화 한다. (L1 정규화)

$$L = \|\vec{\varepsilon}\|^2 + \lambda \sum_{i=0}^K |\beta_i|$$

- Ridge회귀와 마찬가지로 과적합 (overfitting)의 상황이 의심될 때 사용.
- $\lambda$ 가 과하게 크면 편향오류의 증가가 분산오류의 감소를 상쇄하고도 남을 수 있으니 주의한다.
- 회귀계수가 정확하게 0이 될 수 있다.

# I 다항식 회귀

- 다음과 같은 다항식을 사용하여  $X$ 와  $Y$ 사이의 관계를 모형화 한다.

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$$

- 주의할 점은 단 하나의 설명변수  $X$ 가 있다는 것이다.
- 다항식 항은  $I(X^2)$ ,  $I(X^3)$ , 등과 같이 R 수식에 추가한다.

## I 푸아송 회귀

- 종속변수  $Y$ 가 횃수 (count)를 나타내는 경우에 사용한다.

$$\text{Log}(\lambda) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \varepsilon$$

- 푸아송 확률분포함수:

$$P(y) = \frac{\lambda^y e^{-\lambda}}{y!}$$

$$\Rightarrow \text{평균} = \lambda$$

$$\Rightarrow \text{분산} = \lambda$$

$$\Rightarrow \text{표준편차} = \sqrt{\lambda}$$



I 끝.

감사합니다.

