R 교육 세미나

ToBig's 8기 류호성

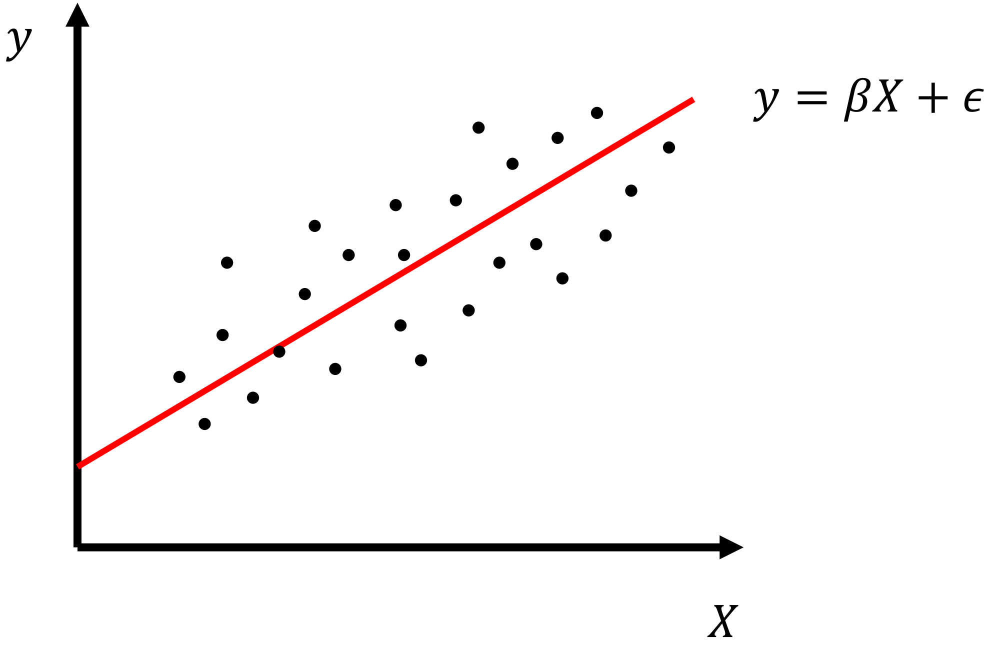
Logistic & Penalized Regression

로지스틱 & 벌점 회귀

Contents

Unit 01 | 회귀분석 Review

회귀분석



회귀식을 잘 찾자!

종속(예측)변수 : 연속형

계수(β)추정 : 최소 제곱 추정

$$S(\beta) = \sum (y - \hat{y})^2$$

변수 선택 : 변수선택기준(AIC , BIC 등)에 부합하는 변수 선택

가정 : 선형성, 비상관성, 정상성, 등분산성, 독립성

고려사항 : outlier / high leverage / 다중공선성

Unit 02 | 그렇다면?

1. 외출시간이 길면 감기에 걸릴까?
2. 수술환자가 살 수 있을까?
3. 아이패드를 사용하는 사람은 아이폰을 사용할까?
4. 질병을 가지고 있는가? (없음, 보균, 양성)
5. 목적지가 멀면 이동수단이 어떻게 되는가? (지하철,버스,택시,도보)

Unit 02 | 그렇다면?

1. 외출시간이 길면 감기에 걸릴까?
2. 수술환자가 살 수 있을까?
3. 아이패드를 사용하는 사람은 아이폰을 사용할까?
4. 질병을 가지고 있는가? (없음, 보균, 양성)
5. 목적지가 멀면 이동수단이 어떻게 되는가? (지하철,버스,택시,도보)

Logistic Regression

로지스틱 회귀는 기존의 회귀 분석과는 다르게 종속 변수가 범주형 데이터를 대상으로 하며 입력 데이터가 주어졌을 때 해당 데이터의 결과가 특정 분류로 나뉘기 때문에 일종의 분류기법으로도 볼 수 있다.

https://en.wikipedia.org/wiki/Logistic_regression

Unit 03 | Logistic regression

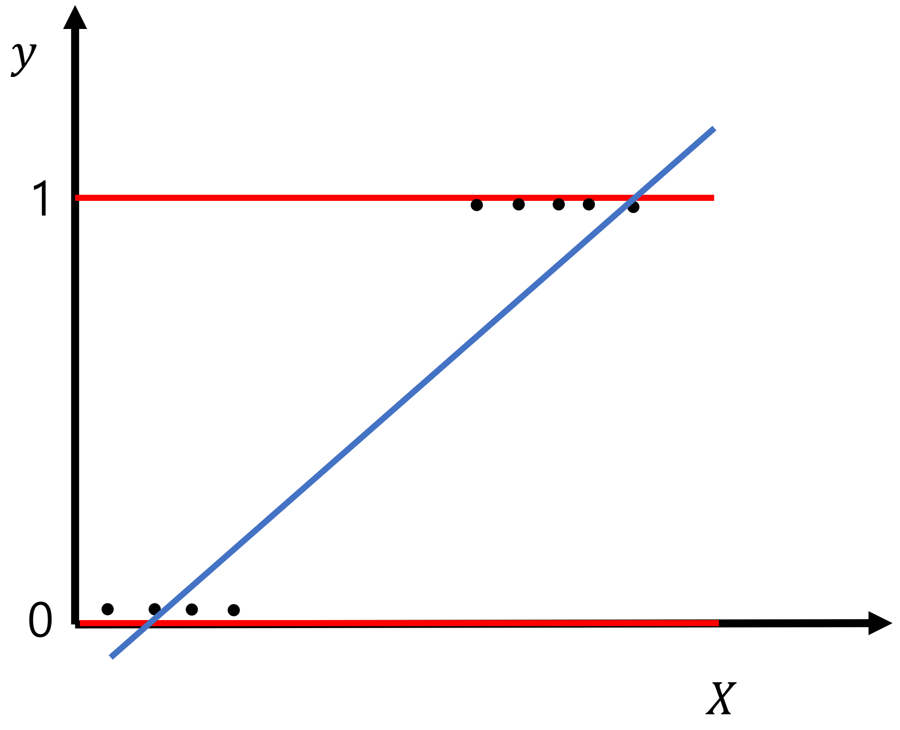
회귀분석에서는

1. 회귀직선식 : $y = \beta X + \epsilon$

2. 계수추정 : $S(\beta) = argmin \sum (y - \hat{y})^2$

이렇게 했었는데 …

Unit 03 | Logistic regression

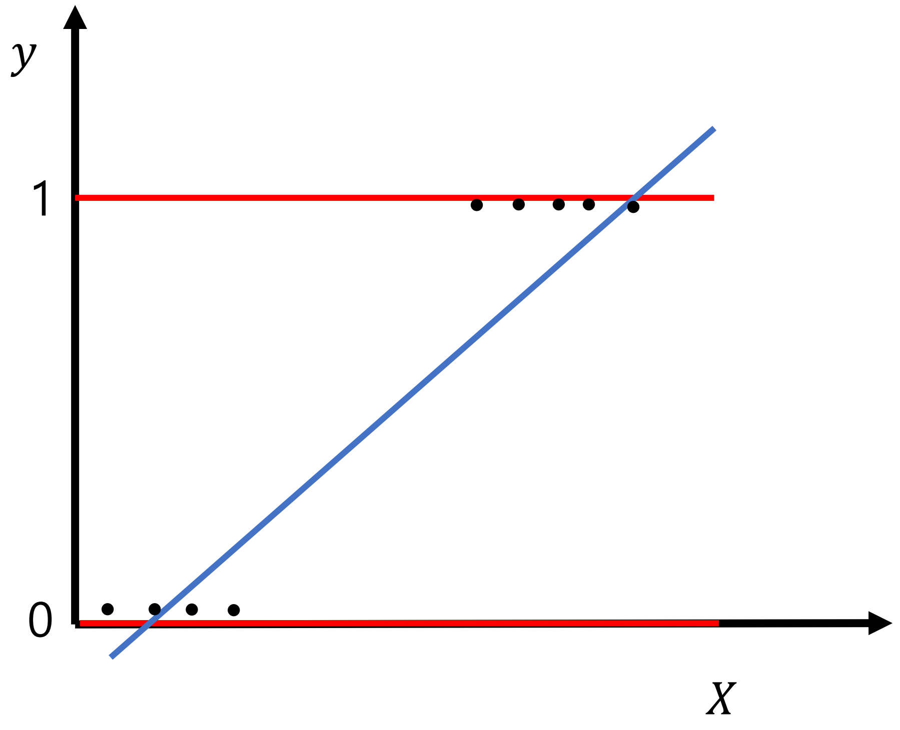


회귀분석 때처럼 해보면..?

1. 회귀직선식 : $y = \beta X + \epsilon$

→ 0보다 작은 값으로 예측되거나, 1보다 큰 값으로 예측이 된다.

Unit 03 | Logistic regression



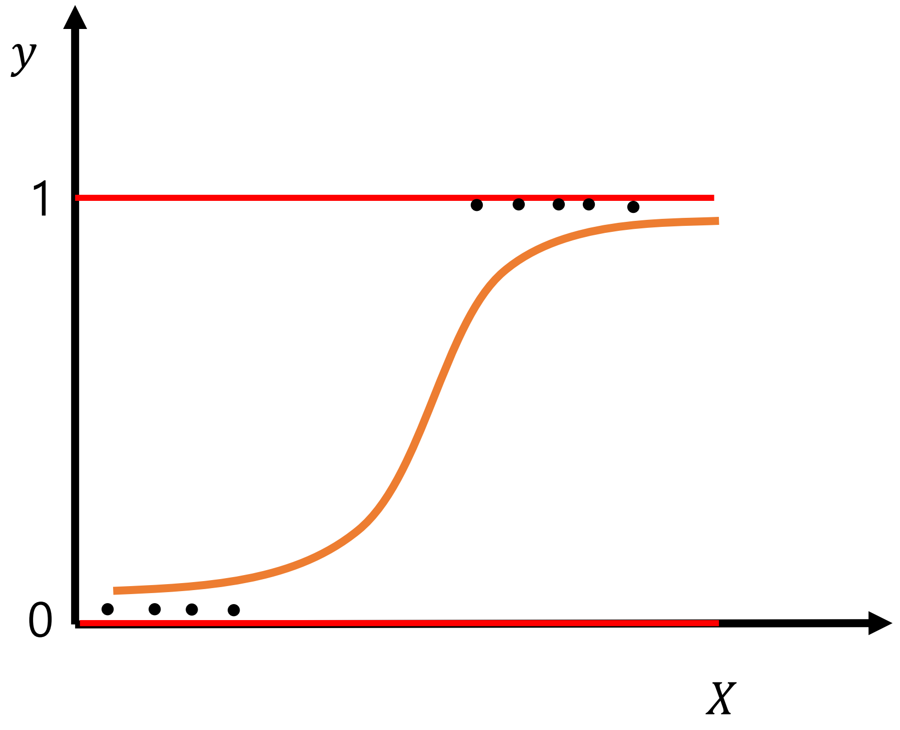
회귀분석 때처럼 해보면..?

1. 회귀직선식 : $y = \beta X + \epsilon$

→ ~~0보다 작은 값으로 예측되거나, 1보다 큰 값으로 예측이 된다.~~

그러면 우리가 예측한 값이 어떻게 돼야 할까?

Unit 03 | Logistic regression



P(Y=1|x) 의 확률을 예측하자!!!!

$$P(Y=1|x) = \frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}}$$

→ 0에서 1로 표현이 된다

Unit 03 | Logistic regression

$$\frac{P(Y=1|x)}{1-P(Y=1|x)} = e^{\beta_0 + \beta_1 x}$$

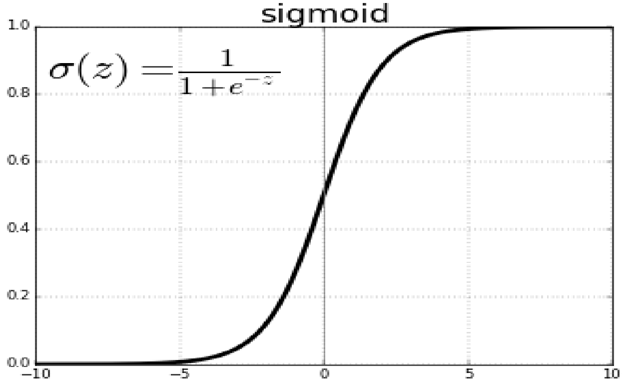
오즈 (Odds) : 실패확률 대비 성공확률
→ 0 ~ Inf 의 값을 갖는다. 이는 $y = e^x$

$$\log\left(\frac{P(Y=1|x)}{1-P(Y=1|x)}\right) = \beta_0 + \beta_1 x$$

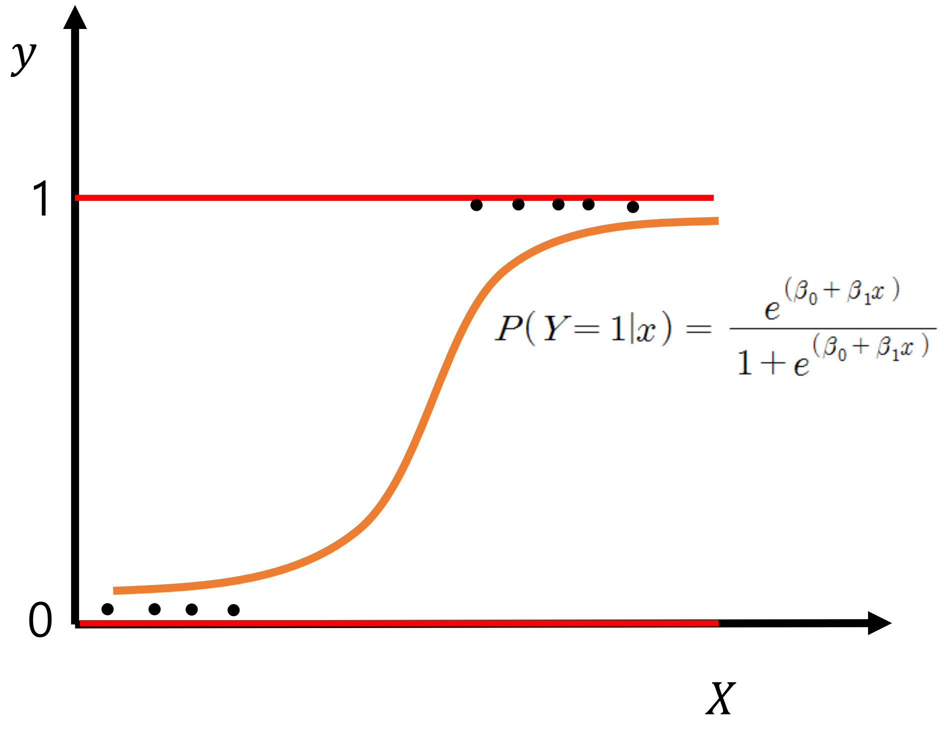
로짓 (Logit) : 선형성을 가진다.
→ -Inf ~ Inf 값을 갖는다.
→ 로지스틱 회귀분석

$$P(Y=1|x) = \frac{e^{(\beta_0 + \beta_1 x)}}{1 + e^{(\beta_0 + \beta_1 x)}}$$

→ 이 식으로 예측!



Unit 03 | Logistic regression



회귀분석에서
계수(β)추정 : 최소 제곱 추정

$$S(\beta) = argmin \sum (y - \hat{y})^2$$

로지스틱 회귀분석에서는 안 된다!

Unit 03 | Logistic regression

→ MLE (Maximum Likelihood Estimation)

$$P(Y=1|x) = \frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}}$$

Likelihood function : $\prod P(x_i)^{y_i}(1-P(x_i))^{1-y_i}$

$$= \boxed{\prod_i P(Y=1|x_i)^{y_i}(1-P(Y=1|x_i))^{1-y_i}}$$

← 얘를 최대화!!하는 β를 찾으면 된다.

→Log 변환 (곱의 연산을 합의 연산으로!) (+→- : 최대 구하는 문제를 최소 구하는 문제로!)

$$\cong -\sum_i y_i \log\big(P(Y=1|x_i)\big) + (1-y_i)\log(1-P(Y=1|x_i))$$

Unit 03 | Logistic regression

예시)

하루 음식 섭취량(x)에 따라 비만(y)이 되는가?
비만 → 1 / not비만 → 0

Train data : x y 값이 있다.
Test data : y 값이 없다.(예측)

Train set (실제 train data의 80%); x/y 모두 주어진다. → 적절한 로지스틱 회귀식을 세운다.

$$\log\left(\frac{P(Y=1|x)}{1-P(Y=1|x)}\right)=\beta_0+\beta_1 x$$

계수 추정 : MLE

Validation set (실제 train data의 20%); 위에서 구한 회귀식을 이용해 y를 예측한다. 실제 y값과 비교!

$$P(Y=1|x)=\frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}}$$

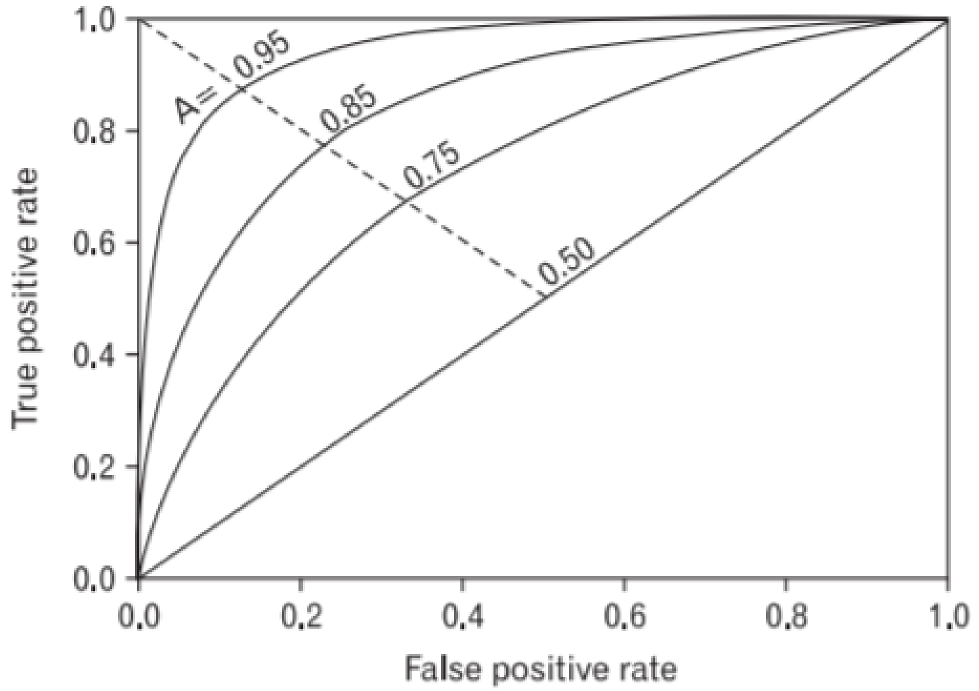
> 0.5 → 1 아니면 0으로 예측

Test set ; x가 주어지고, 우리가 만든 로지스틱 회귀식으로 분류예측을 한다.

Unit 03 | Logistic regression

내가 세운 모델이 잘된 모델인가?? → ROC 커브 곡선 확인

		실제		총
		비만(1)	Not 비만(0)	
예측	비만(1)	4012	987	4999
	Not비만 (0)	1920	13045	14965
총		5932	14032	19964



Accuracy : 0.8543
민감도 (True Positive) : 0.6763
특이도 (True Negative) : 0.9296

Unit 04 | Multinomial Logistic Regression

설명변수 1개 / 종속변수 Binomial 일 때

$$\log\left(\frac{P(Y=1|x)}{1-P(Y=1|x)}\right)=\beta_0+\beta_1 x$$

설명변수 p개 / 종속변수 Binomial 일 때

$$\log\left(\frac{P(Y=1|x)}{1-P(Y=1|x)}\right)=\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\ldots.+\beta_p x_p$$

→ 종속변수의 class가 여러 개인 Multinomial 일 때는?

Unit 04 | Multinomial Logistic Regression

5. 목적지가 멀면 이동수단이 어떻게 되는가? (지하철,버스,택시,도보)

→ 다수의 binomial logistic regression 결합으로 풀면 된다!

$$P(Y = subway|x) = \frac{e^{\beta_{subway}X}}{1 + e^{\beta_{subway}X}}$$

지하철인지 아닌지

$$P(Y = bus|x) = \frac{e^{\beta_{bus}X}}{1 + e^{\beta_{bus}X}}$$

버스인지 아닌지

$$P(Y = taxi|x) = \frac{e^{\beta_{taxi}X}}{1 + e^{\beta_{taxi}X}}$$

택시인지 아닌지

$$P(Y = walk|x) = \frac{e^{\beta_{walk}X}}{1 + e^{\beta_{walk}X}}$$

도보인지 아닌지

Unit 04 | Multinomial Logistic Regression

좋은 아이디어지만 독립적으로 하면 복잡하다!
그래서 쪼~금 변형을 한다.

→ softmax

softmax : $P(Y = k|x) = \frac{e^{\beta_k X}}{\sum_k e^{\beta_k X}}$

→ 이 값이 최대가 되는 k 일 때로 예측

https://www.youtube.com/watch?v=MFAnsx1y9ZI

Unit 04 | Multinomial Logistic Regression

그렇다면 계수 추정할 때는?

→ Cross – entropy

$$\text{Cross-entropy} : -\sum_i \sum_k 1\{y_i = k\} * \log\left(\frac{e^{\beta_k X_i}}{\sum_k e^{\beta_k X_i}} \right)$$

*** $1\{y_i = k\}$: y_i가 k 일 때 1 아니면 0

이 값을 최소화하는 β 를 찾는다!

https://www.youtube.com/watch?v=jMU9G5WEtBc

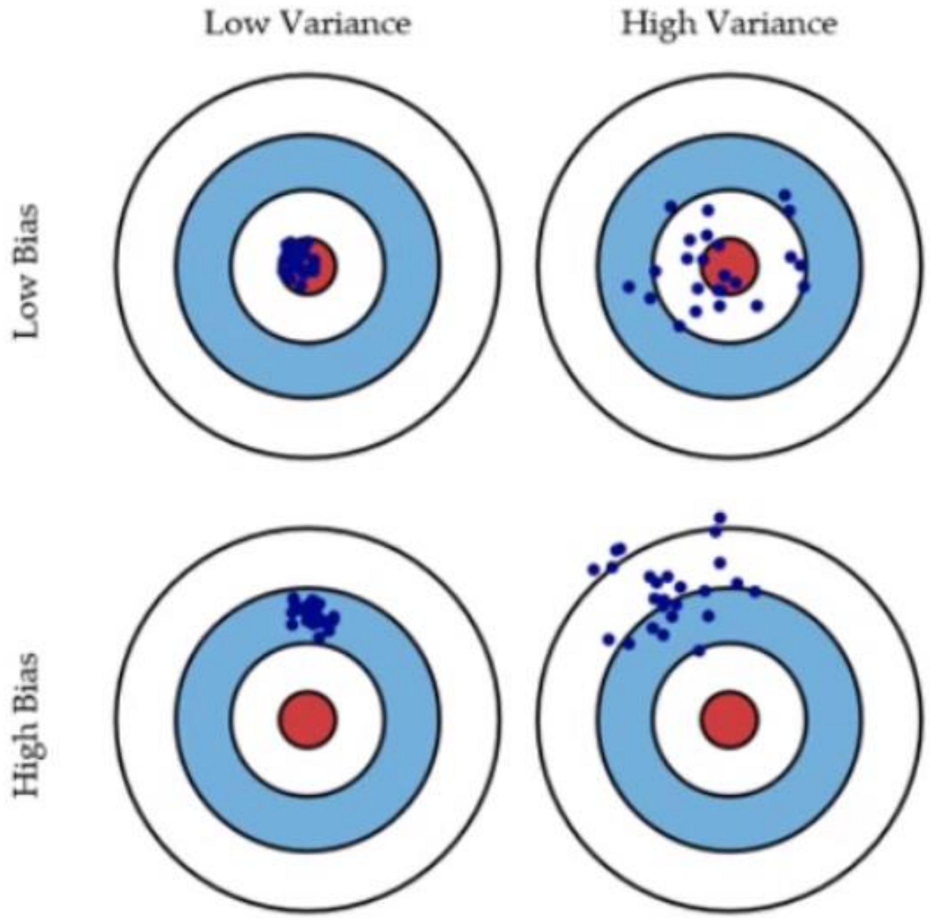
Unit 05 | Penalized regression

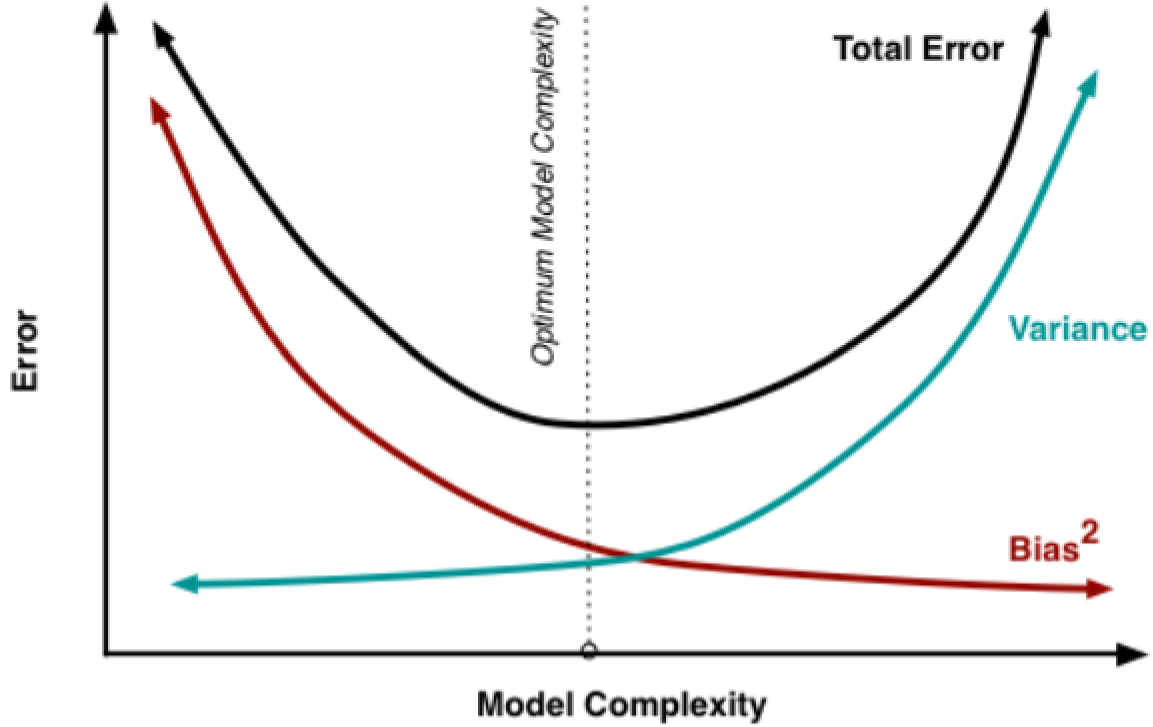
여기서부터는 회귀와 로지스틱 회귀 모두에게 해당하는 내용!

$$\mathrm{E}\left[\left(y-\hat{f}(x)\right)^2\right]=\mathrm{Bias}\left[\hat{f}(x)\right]^2+\mathrm{Var}\left[\hat{f}(x)\right]+\sigma^2$$

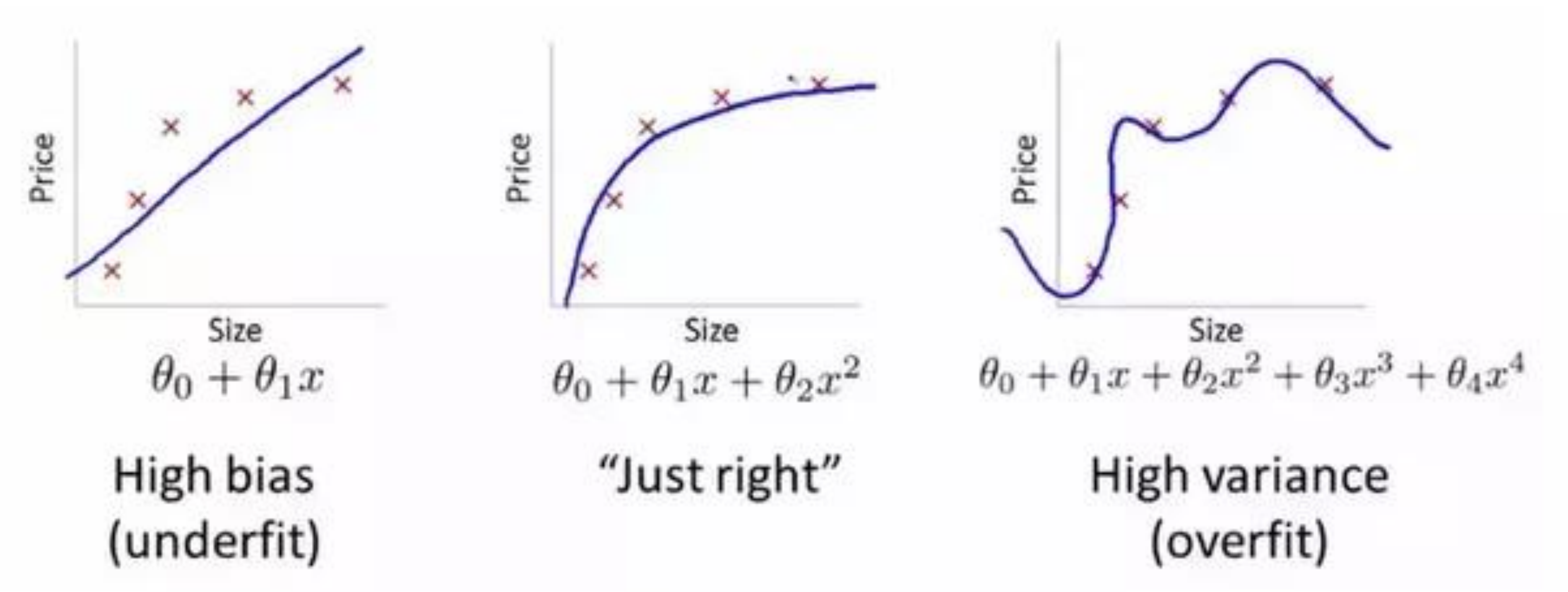
편향(bias) & 분산(Variance) 간의 Trade-off

Unit 05 | Penalized regression





Unit 05 | Penalized regression



Unit 05 | Penalized regression

Under fitting (high bias, low variance)일 때 → 모형에 새로운 변수를 추가하면 된다.

Over fitting (low bias, high variance) 일 때

→

1. regularization
2. 데이터 수 늘리기
3. shrinkage 방법 (모델 복잡도 줄이기)
   - ridge regression
   - lasso regression

능형회귀 (ridge regression)

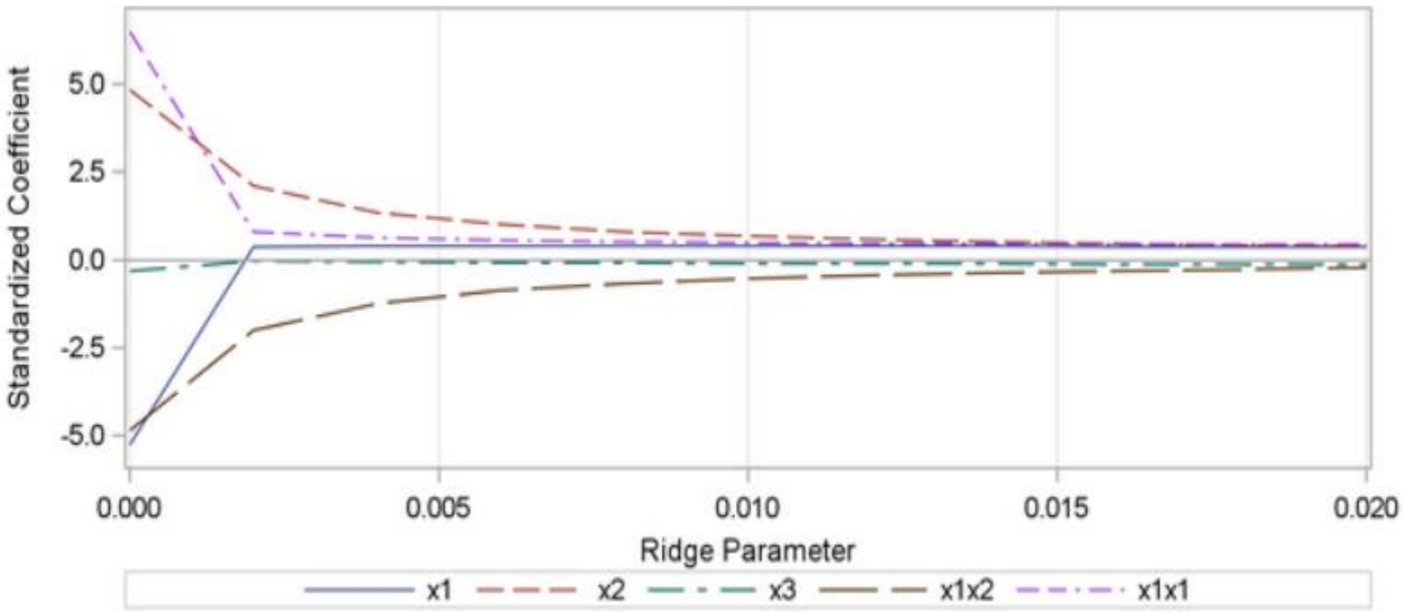
: 설명 변수(p개) >> 데이터 수(n개) 일 때,
변수 간의 연관관계(다중 공선성)가 심한 경우
→ 계수를 추정할 때, penalized term(L2) 를 추가해 변수의 영향력을 줄인다.

$$\beta^{ridge} = \mathrm{argmin}_{\beta}\left\{\sum_{i}^{n}(y - BX)^2 + \lambda\sum_{j}^{p}\beta_j^2\right\}$$

적절한 λ(하이퍼 파라미터)를 찾는 것이 관건!

Unit 05 | Penalized regression

적절한 Lambda(λ) 찾기



Lambda(λ) 의 변화에 따른 Beta(β) 계수의 축소를 보여주는 Graph

→ 기울기가 안정되어 가는 지점의 λ를 선택!

Beta(β) 계수들이 0에 가까워지고 있기는 하나, 0은 아니다.

→ 영향력은 줄었지만, 아직 p개의 모든 변수를 사용한다.

Unit 05 | Penalized regression

Lasso regression

: ridge 회귀와 비슷한 역할을 하지만, Beta(β) 계수를 0으로 만드는 점에서 차이가 있다.

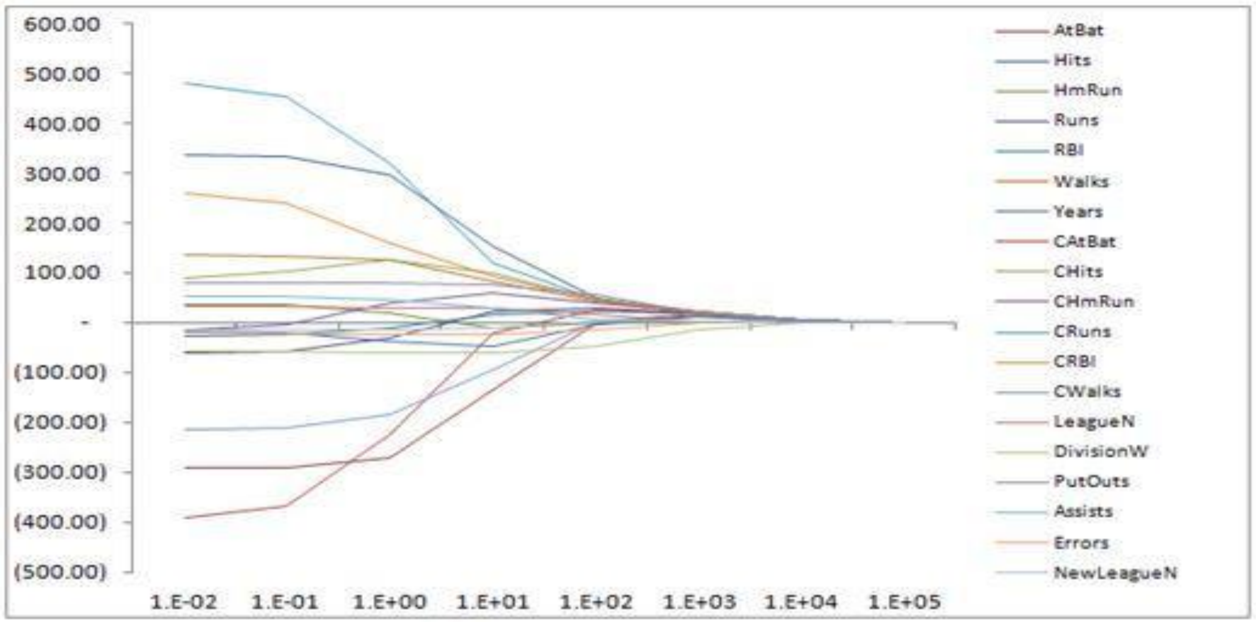
→ 변수 선택의 역할이 생겼다.

→ 계수를 추정할 때 penalized term(L1)을 사용한다

$$\beta^{lasso} = \mathrm{argmin}_{\beta}\left\{\sum_{i}^{n}(y - BX)^2 + \lambda\sum_{j}^{p}|\beta_j|\right\}$$

마찬가지로, 적절한 λ 찾기!

적절한 Lambda(λ) 찾기



Lambda(λ) 의 변화에 따른 Beta(β) 계수가 완전히 0이 되는 지점이 생긴다.

→ 변수 선택 효과

https://brunch.co.kr/@itschloe1/11

Q & A

들어주셔서 감사합니다.