16기 분석 2주차 세션
Regression

15기 분석 최은지

Machine Learning

머신 러닝
Machine Learning

지도 학습
Supervised Learning

- 선형회귀, 논리회귀, 의사결정트리 등-

비지도 학습
Unsupervised Learning

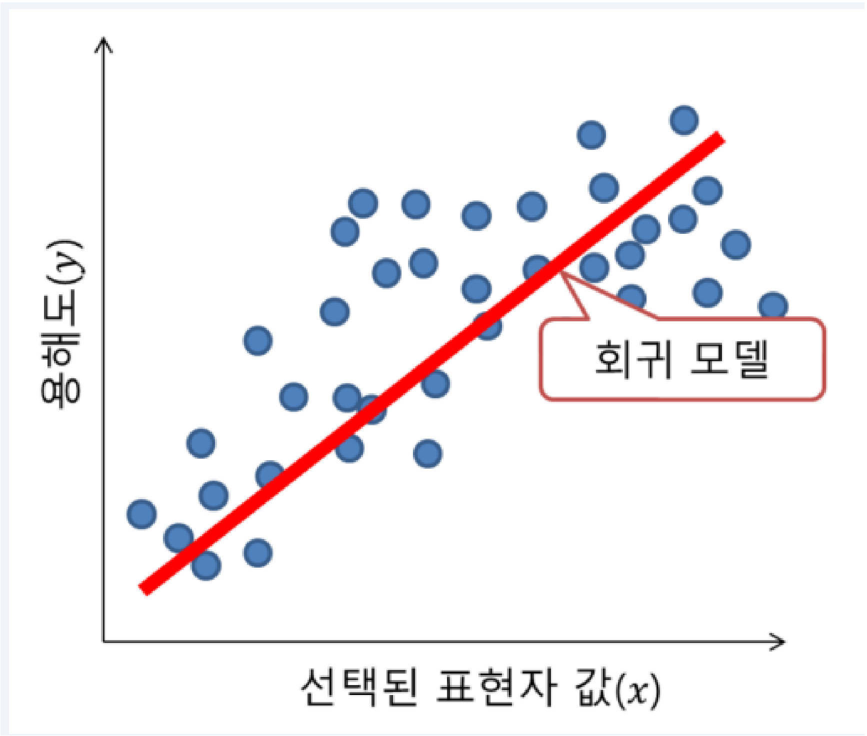
- 클러스터링 등 -

강화 학습
Supervised Learning

Supervised Learning

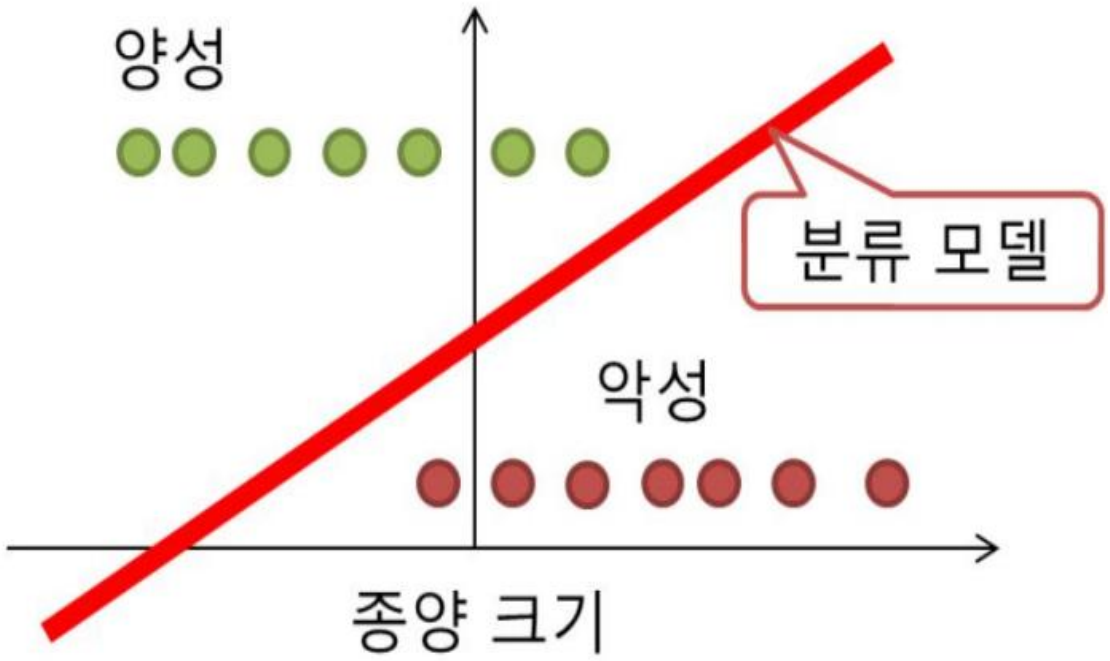
종속 변수의 형태가 무엇이냐에 따라!

1. 회귀(연속형)



집값 예측, GDP 예측 등

2. 분류(범주형)



스팸 분류기, 악성종양 판별 등

Linear Regression

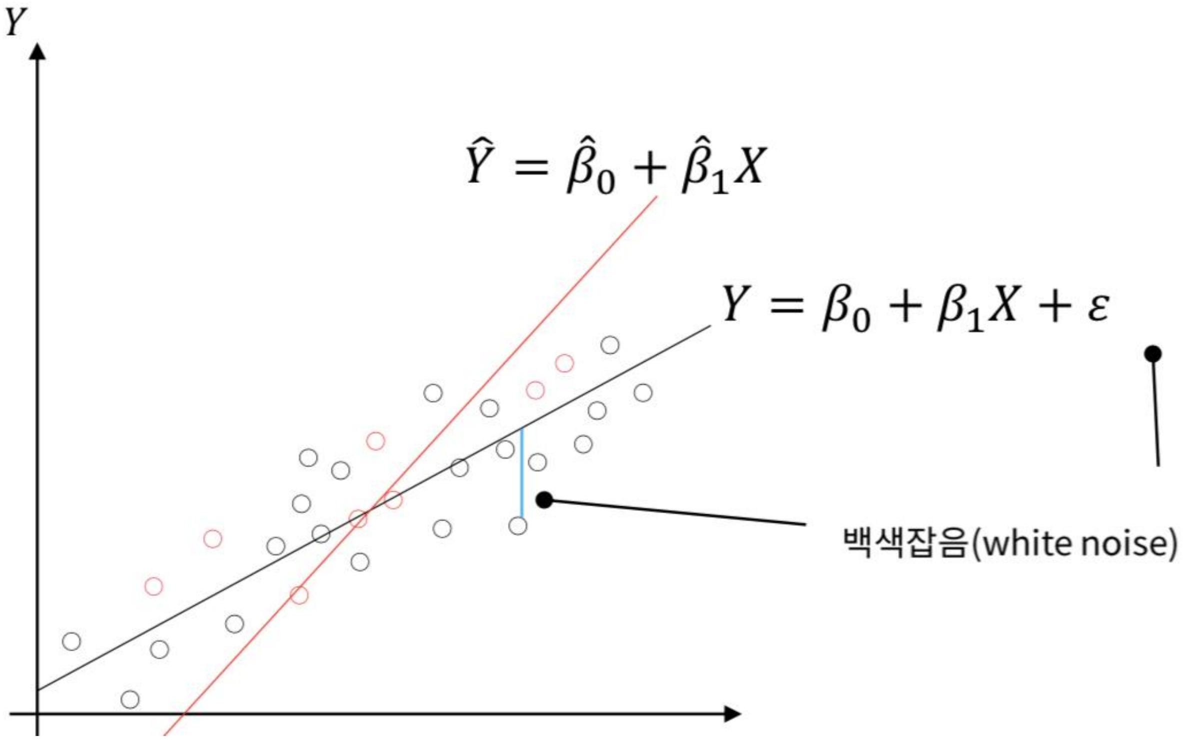
1-1. 단순선형회귀 (입력 변수 X가 1개일 때)

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

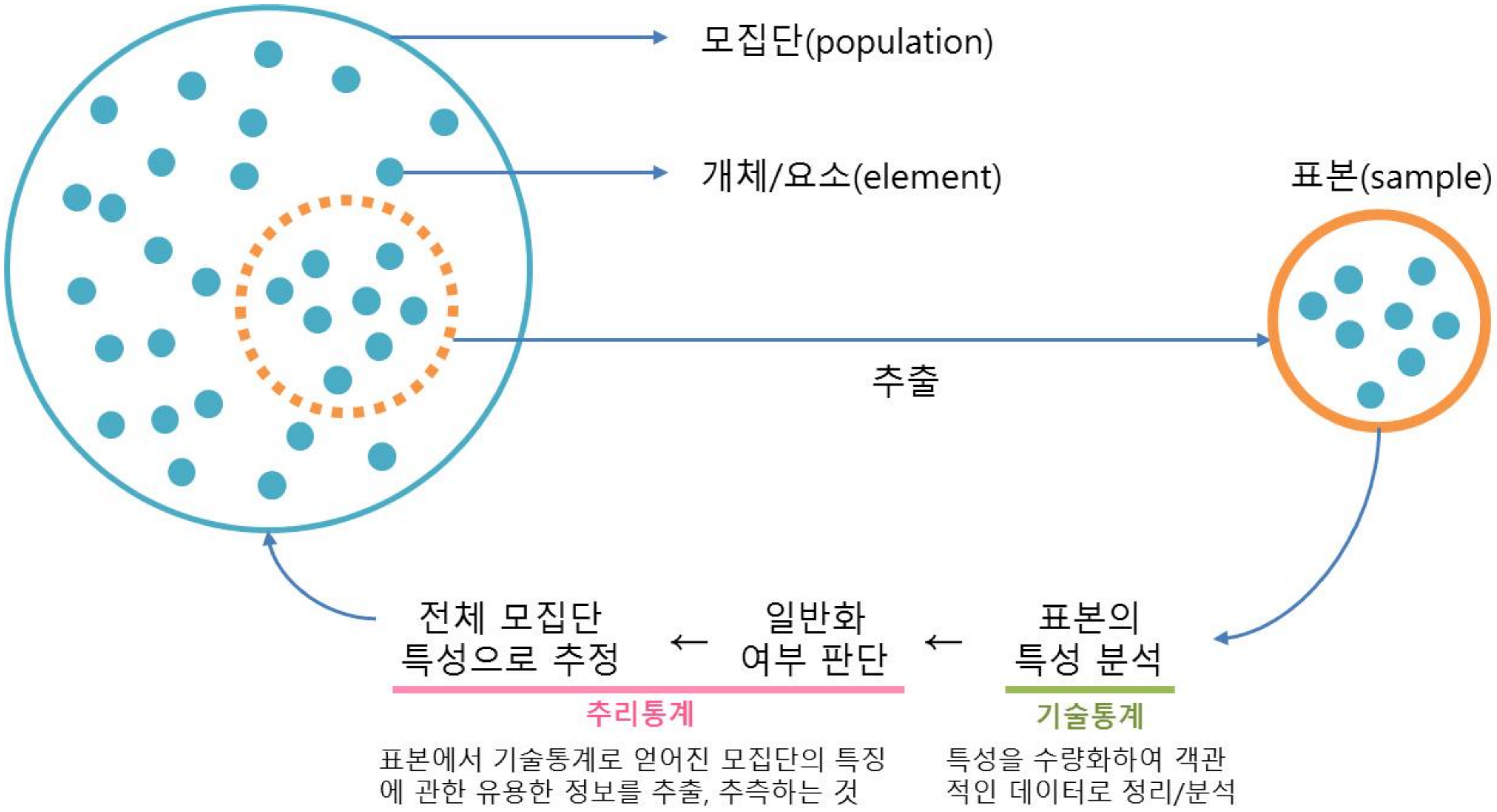
모집단 모수(β)를 모르니
표본으로 추정해야 해서

추정회귀식

$$\hat{y} = b_0 + b_1 x$$

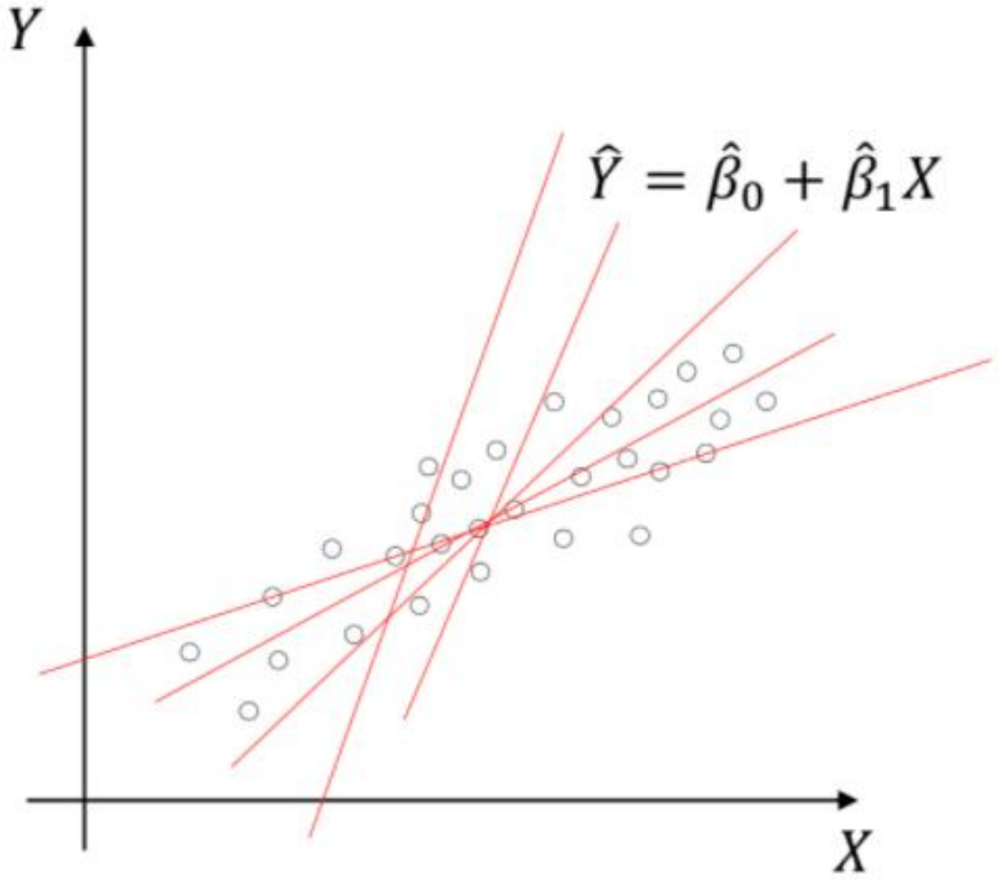


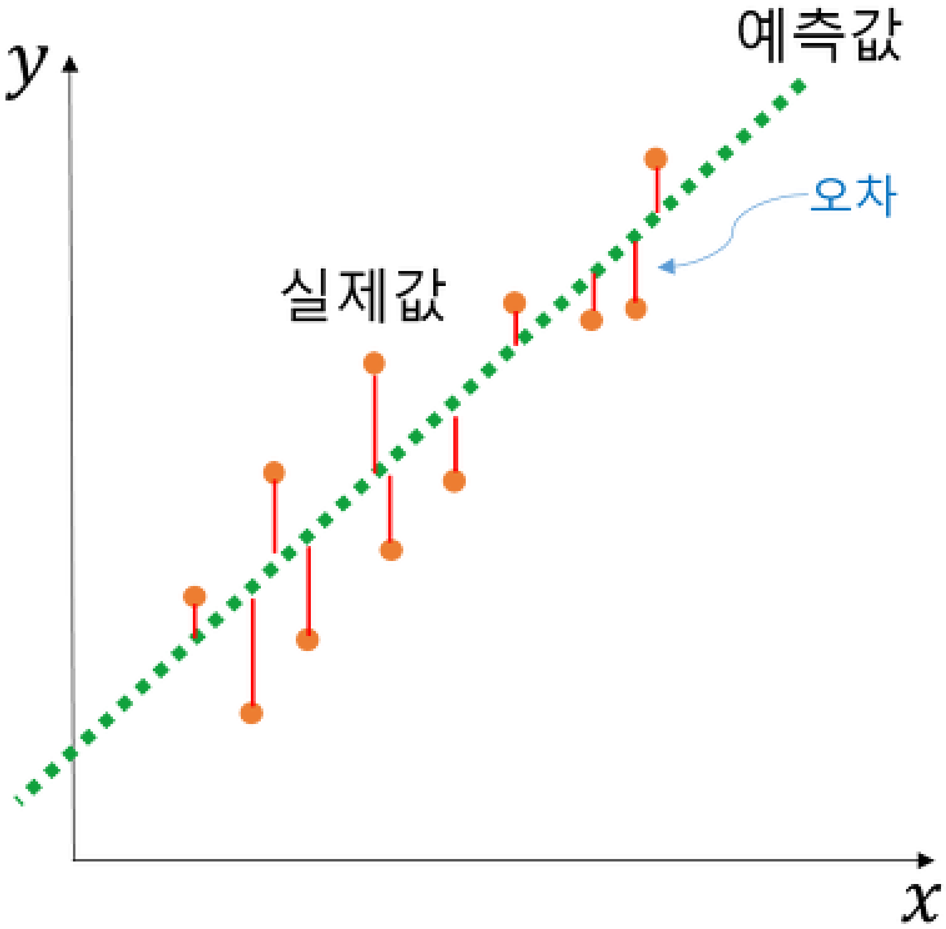
Linear Regression



Linear Regression

수많은 직선 중 무엇을 골라야 할까?

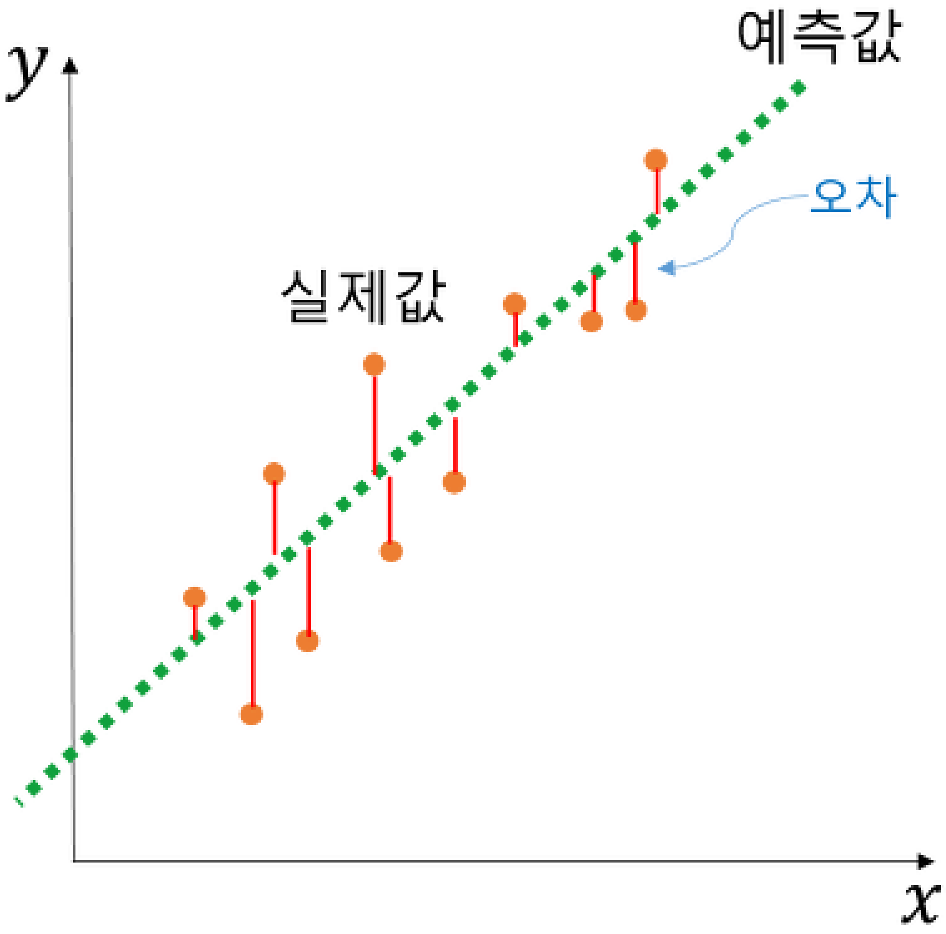




잔차를 최소화하는 직선!

Linear Regression

잔차의 제곱합 (SSE : Error Sum of Squares)



$$e_i = y_i - \hat{y}_i = y_i - (\widehat{\beta_0} + \widehat{\beta_1} x_i)$$

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - (\widehat{\beta_0} + \widehat{\beta_1} x_i)\right)^2$$

Q. 왜 잔차를 제곱하여 사용할까?

1. 잔차의 합이 0이 될 수 있기 때문

2. 미분이 가능해야하기 때문

Linear Regression

미분을 통한 극소값(오차가 가장 작은 지점) 찾기

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (\widehat{\beta_0} + \widehat{\beta_1} x_i))^2$$ = Lost function (L)

$$\frac{\partial L}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial L}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0$$

Linear Regression

미분을 통한 극소값(오차가 가장 작은 지점) 찾기

$$\frac{\partial L}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial L}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0$$

$$\sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i - n\beta_0 = 0 \longrightarrow \beta_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i \longrightarrow \therefore \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} {x_i}^2 = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i - \left(\frac{1}{n}\sum_{i=1}^{n} y_i - \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i\right)\sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} {x_i}^2 = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i + \beta_1 \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 - \beta_1 \sum_{i=1}^{n} {x_i}^2 = 0$$

$$\Rightarrow \beta_1 \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 - \beta_1 \sum_{i=1}^{n} {x_i}^2 = \frac{1}{n}\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i$$

$$\therefore \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Linear Regression

선형 회귀 정확도 평가

1. MSE (자유도 n-2)

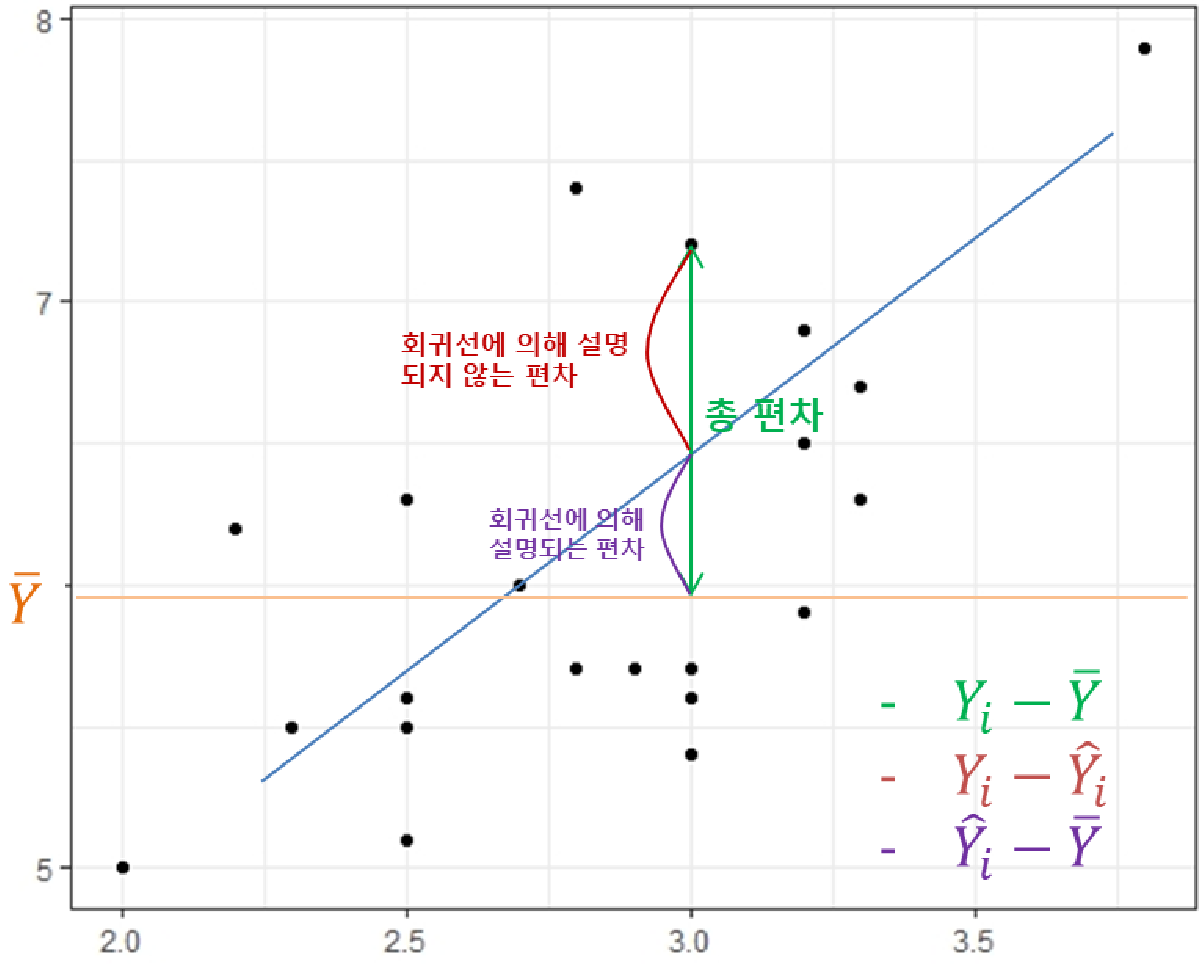
$$\mathrm{MSE} = \frac{SSE}{n-2} = \frac{1}{n-2}\Sigma(\hat{y} - y)^2$$

2. RMSE

$$\mathrm{RMSE} = \sqrt{MSE} = \sqrt{\frac{\Sigma(\hat{y}-y)^2}{n}}$$

Linear Regression

3. R - Squared



$$y_i - \bar{y} \;=\; y_i - \hat{y}_i \;+\; \hat{y}_i - \bar{y}$$

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{SST (Total sum of squares)}} = \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{SSE (Error sum of squares)}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{SSR (Regression sum of squares)}}$$

입력 변수로 설명할 수 없는 변동 비율

$$R^2 = \frac{SST - SSR}{SST} = 1 - \frac{SSE}{SST} = \frac{SSR}{SST} \qquad where\ SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

=> 1에 가까울 수록 좋다!

Linear Regression

1-2. 다중선형회귀 (입력 변수 X가 여러 개일 때)

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$$

[회귀 계수의 추정 – 편미분]

$$\frac{\partial L}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}) = 0$$

$$\frac{\partial L}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})x_{1i} = 0$$

$$\vdots \qquad\qquad \vdots$$

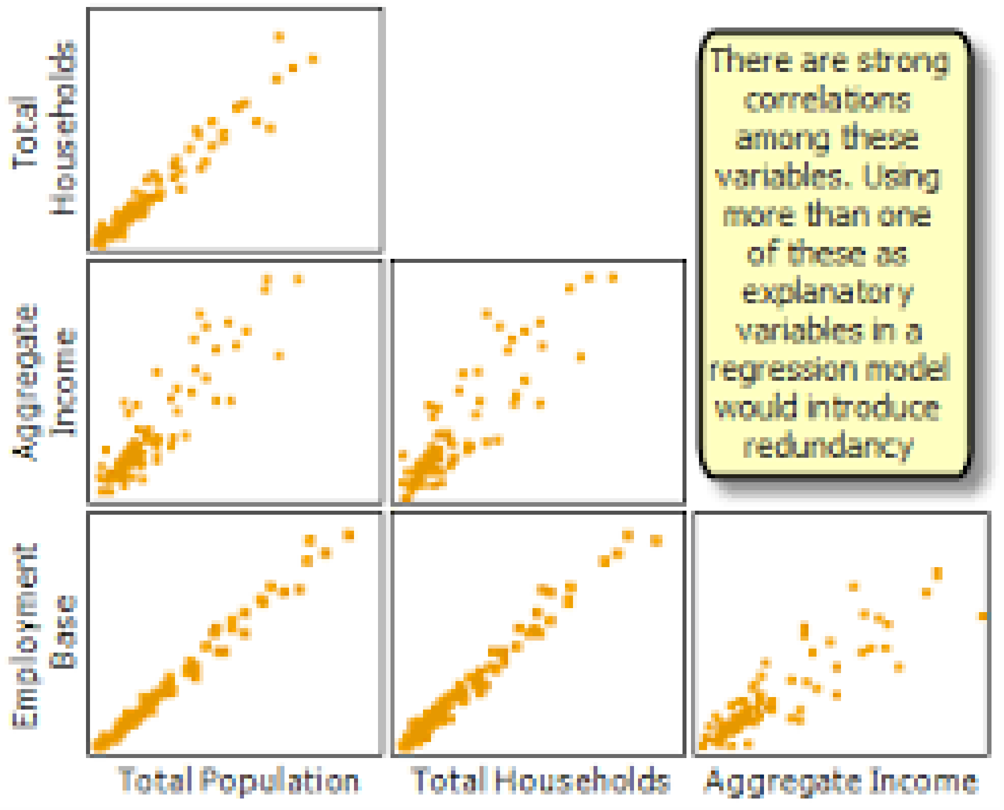
$$\frac{\partial L}{\partial \beta_p} = -2\sum_{i=1}^{n}(y_i - \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})x_{pi} = 0$$

Linear Regression

다중공선성 (Multicollinearity) 문제!

-> 일부 독립변수가 다른 독립 변수와 높은 상관관계를 가질 때 발생하는 문제

-> 회귀선의 판단 능력 저하



Linear Regression

VIF (Variance inflation factor) : 10이상인 경우 다중공선성이 있다고 판단

$$\mathrm{VIF_i} = \frac{1}{1 - R_i^2}$$

R_i^2 이 크다!

-> Xi를 제외한 다른 X 변수들이 Xi를 잘 설명한다!

-> 변수간 관계가 높을수록 VIF 값은 커진다!

$$x_1 = \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_p X_p + \varepsilon$$

$$\downarrow$$

$$R_1{}^2$$

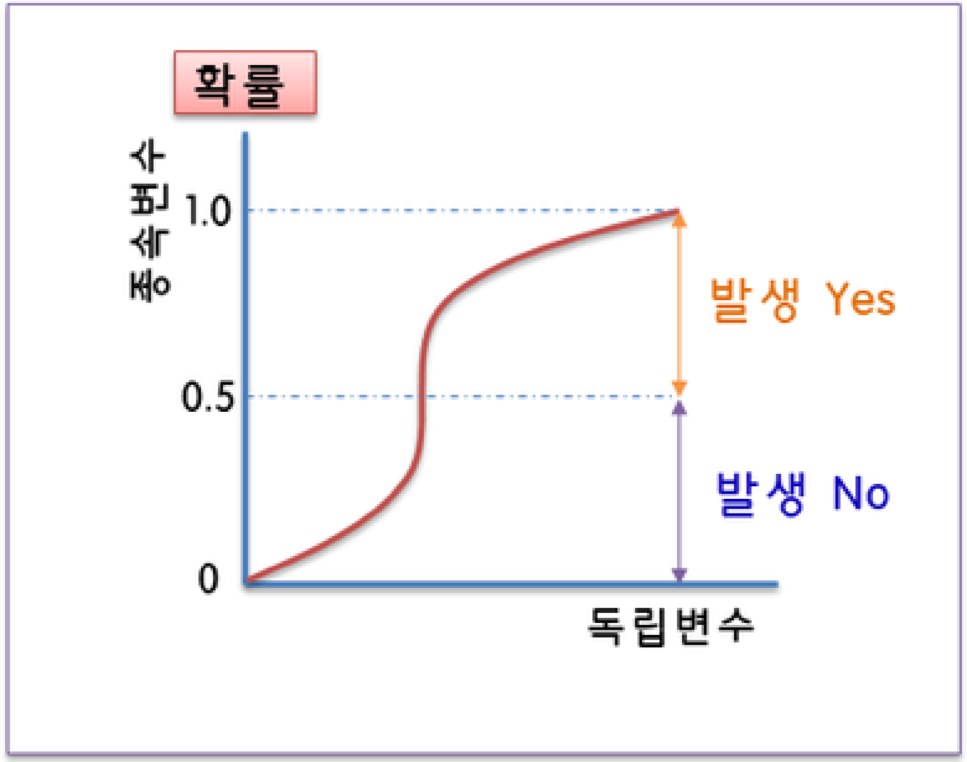
$$\downarrow$$

$$VIF_1 = \frac{1}{1-R_1{}^2}$$

$R^2 > 0.9$ 이상인 경우, $\mathrm{VIF} > 10$

Logistic Regression

2. 로지스틱 회귀



로지스틱 회귀분석

- 종속변수가 범주형일 때

- 0/1, 합격/불합격, 사망/생존, poor/good/excellent, count(하루에 마시는 물이 몇 잔인지 등)

- 확률 값을 구하고 label을 예측

Logistic Regression

2. 로지스틱 회귀

[기존의 선형회귀식]

$$p(X) = P(success|X_1, \cdots, X_k) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + randome\ error(\varepsilon)$$

$$\downarrow$$

$$ln(p(X)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$$

$$\downarrow$$

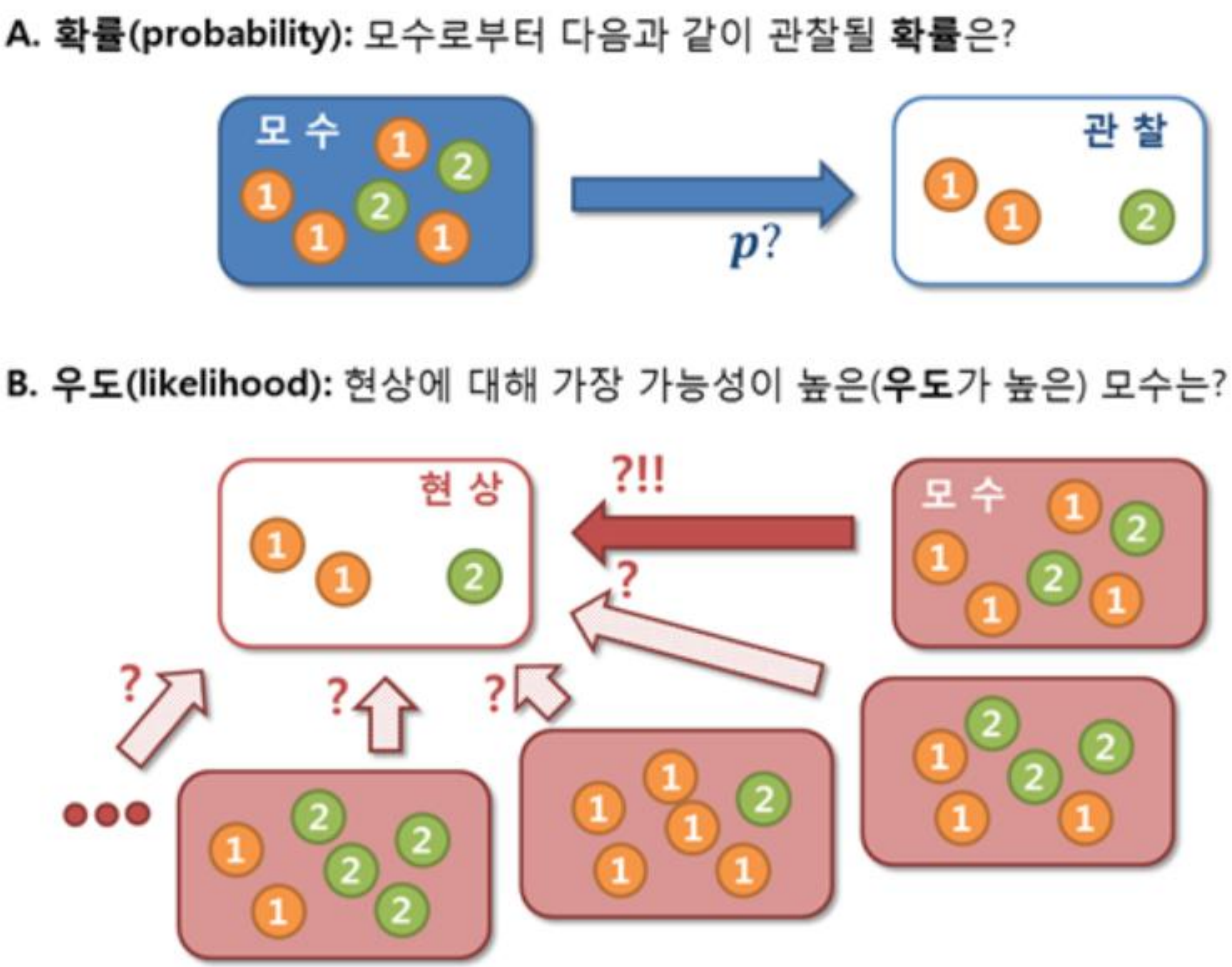
$$logit = ln\left(\boxed{\frac{p(X)}{1 - p(X)}}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$$

$$\downarrow$$

$$Y = p(X) = \frac{e^{\beta_0 + \beta_1 X + \cdots + \beta_k X_k}}{1 + e^{\beta_0 + \beta_1 X + \cdots + \beta_k X_k}}$$

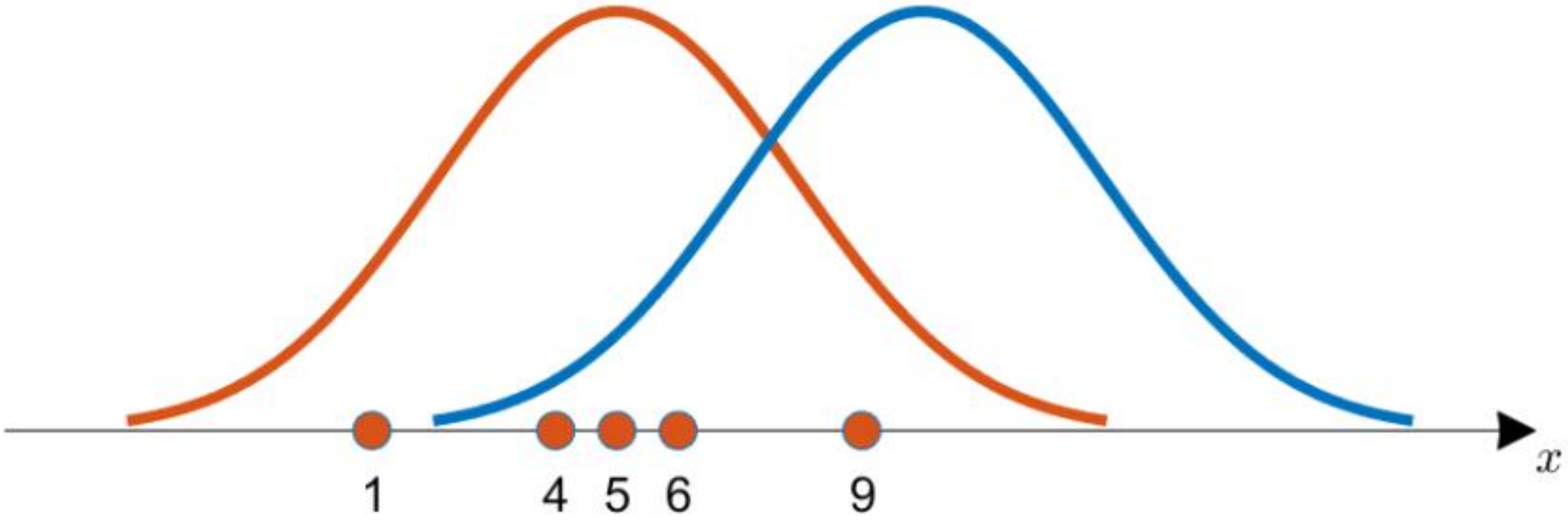
Logistic Regression

Likelihood



Logistic Regression

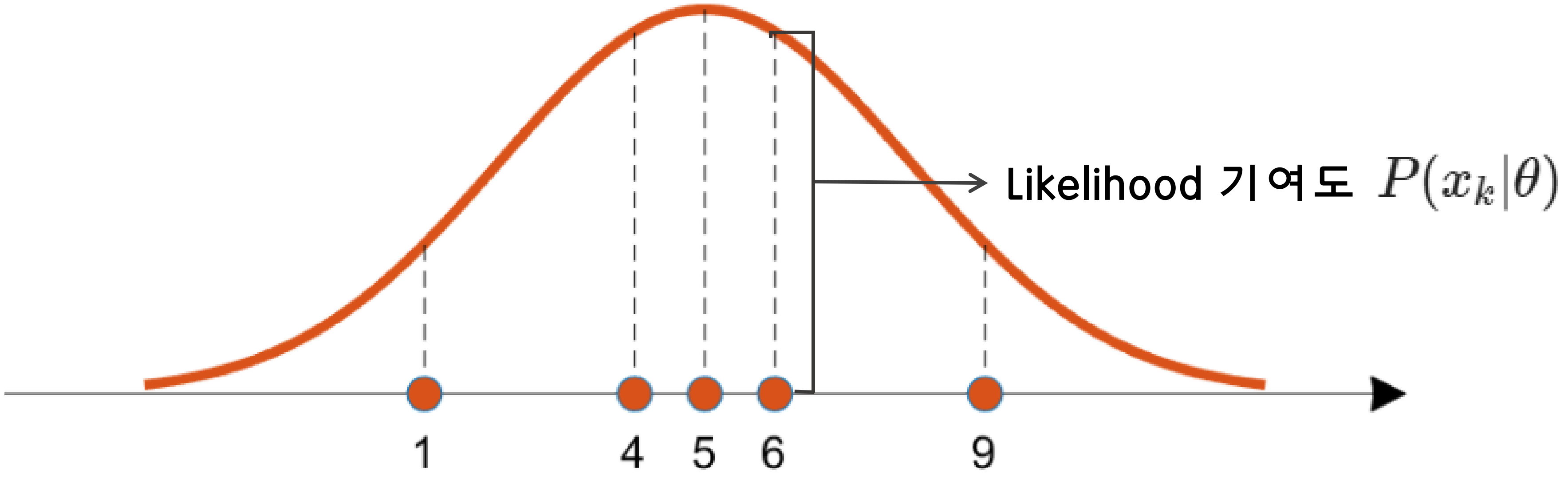
주어진 데이터는 어떤 분포로부터 추출되었을까?



Likelihood(가능도) : 관측 값이 어떤 분포에 해당할 확률

Logistic Regression

주어진 데이터는 어떤 분포로부터 추출되었을까?



Likelihood function : 전체 표본집합의 결합확률밀도 함수

$$P(x|\theta) = \prod_{k=1}^{n} P(x_k|\theta)$$

$$L(\theta|x) = \log P(x|\theta) = \sum_{i=1}^{n} \log P(x_i|\theta)$$

$$\frac{\partial}{\partial\theta} L(\theta|x) = \frac{\partial}{\partial\theta} \log P(x|\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log P(x_i|\theta) = 0$$

Logistic Regression

[베르누이 시행]

- 결과가 두 가지 중
  하나로만 나오는 실험이나 시행

$$X \sim \mathrm{Bern}(x;\mu)$$

$$\mathrm{Bern}(x;\mu) = \begin{cases} \mu & \text{if } x = 1, \\ 1-\mu & \text{if } x = 0 \end{cases}$$

[베르누이 확률분포]

$$\mathrm{Bern}(x;\mu) = \mu^{x}(1-\mu)^{(1-x)}$$

[회귀계수의 추정]

$$f(y_i;\pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

⬇

$$L(\beta|X,y) = f(y_1;\pi_1)\cdot\ \cdots\ \cdot f(y_n;\pi_n) = \prod_{i=1}^{n}\pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

⬇

$$l(\beta|X,y) = \sum_{i=1}^{n}\{y_i log\pi_i + (1-y_i)log(1-\pi_i)\}$$

⬇

이 때, $\pi_i = \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}$

⬇

$$\frac{\partial l(\beta)}{\partial \beta} = \sum x_i(y_i - \pi_i) = 0$$

Logistic Regression

[회 귀 계 수 의 해 석]

Logit : RF_impedance가 1단위 증가할때 불량일 logit이 -0.0468단위 증가한다.

Odds : RF_impedance가 1단위 증가할때 불량일 확률이 0.954배(exp(-0.0468)) 증가한다

$$odds = \frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$$

도박에서
승률을 의미

$$logit = \log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

회귀식을 선형으로
변환하는 함수

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
Pressure	0.0057	0.0002	24.74	<0.0001
CL2 Flow	0.0030	0.0082	0.37	0.7115
RF_impedance	−0.0468	0.0172	−2.74	0.0062

Logistic Regression

[분류 평가 지표]

실제 클래스 (Actual Class) \ 예측 클래스 (Predicted Class)	Negative(0)	Positive(1)
Negative(0)	TN (True Negative)	FP (False Positive)
Positive(1)	FN (False Negative)	TP (True Positive)

$$Accuracy = \frac{TN + TP}{TN + FP + FN + TP}$$

Logistic Regression

[분류 평가 지표]

A 병원 (잘된 분류 / 잘못된 분류)

실제값	예측값	
	암환자	일반환자
암환자	9	1
일반환자	30	60

B 병원 (잘된 분류 / 잘못된 분류)

실제값	예측값	
	암환자	일반환자
암환자	1	9
일반환자	20	70

[정밀도 (Precision)]

정밀도 = TP/(TP+FP)

A병원 = 9/(9+30) =0.23
B병원 = 1/(1+20) = 0.04

[재현율 (Recall)]

재현율 = TP / (TP+FN)

A병원 = 9/(9+1) =0.9
B병원 = 1/(1+9) = 0.1

[F1 점수 (F1-score)]

F1점수 = 2*재현율*정밀도/(재현율+정밀도)

참고 문헌

[회귀 관련 자료]

Regression(02) - 다중선형회귀 및 다중공선성 | DataLatte's IT Blog (heung-bae-lee.github.io)

[MLE]

최대우도법(MLE) - 공돌이의 수학정리노트 (angeloyeo.github.io)
Logistic regression (tistory.com)

[혼동 행렬]
혼동행렬 / 정확도 / 정밀도 / 재현율 / F1 점수 (tistory.com)

감사합니다!