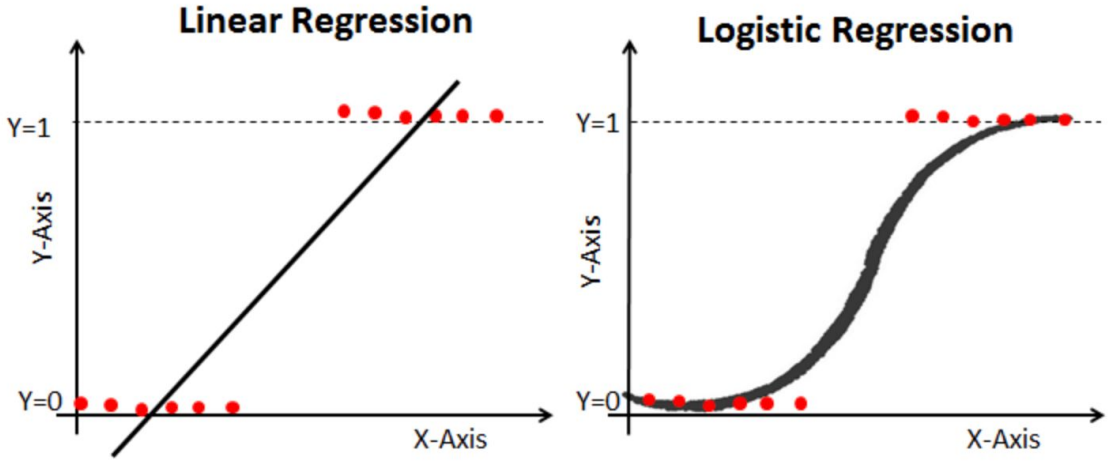
# 로지스틱 회귀

# Logistic Regression

# 로지스틱 회귀분석

- 선형회귀분석에서는 종속변수(y)가 연속형
- 로지스틱 회귀분석은 종속변수가 범주형이며 0 or 1인 경우 사용
  - 머신러닝에서 이진분류(Binary Classification) 모델로 사용
  - 데이터가 어떤 범주에 속할 확률을 0에서 1 사이의 값으로 예측하고 그 확률에 따라 가능성이 더 높은 범주에 속하는 것으로 분류해주는 지도 학습 알고리즘
  - ex) 대학합격여부, 암 재발여부, 상품 구매여부, 보험가입여부 등

# 로짓(logit)변환과 오즈(odds)




- y가 0 또는 1(성공/실패, 생존/사망, 합격/불합격 등)인 경우에는 왼쪽 차트처럼 선형회귀로 fitting하기 어려움(잘 적합하지 않으며, y의 추정값이 0보다 작거나 1보다 커질수도 있음)
- 곡선으로 fitting 하기 위해 로짓함수(시그모이드 함수)를 사용

# 로짓(logit)변환과 오즈(odds)

- 오즈(odds) = 실패에 비해 성공할 확률의 비율 = p / (1-p)
- 게임에서 이길 확률이 ⅕, 질 확률이 1-⅕ = ⅘ 라면, 게임에 이길 오즈는 ¼
  - 게임을 5번 한다면 4번 지는 동안 1번 이긴다 라고 해석
- 로짓(logit) = log 오즈 = log(p/(1-p))

# 로짓(logit)변환과 오즈(odds)

| 구분 | a약 | b약 | 전체 |
|---|---|---|---|
| 사망 | 32 | 24 | 56 |
| 생존 | 20 | 42 | 62 |
| 전체 | 52 | 66 | 118 |

## odds(A)

- P(A) = 20/52 = 0.38
- 1-P(A) = 0.62
- odds(A) = 0.38 / 0.62 = 0.61
- A 약을 먹으면 100명 사망할 동안 61명 생존

## odds(B)

- P(B) = 42/66 = 0.63
- 1-P(B) = 0.37
- odds(B) = 0.63 / 0.37 = 1.7
- B 약을 먹으면 100명 사망할 동안 170명 생존

## Odds Ratio(오즈비)

- B에 대한 A의 오즈비 = 0.61 / 1.7 = 0.36
- B에 비해 A 일 때, 생존이 0.36배 = 64%가 생존율이 떨어짐

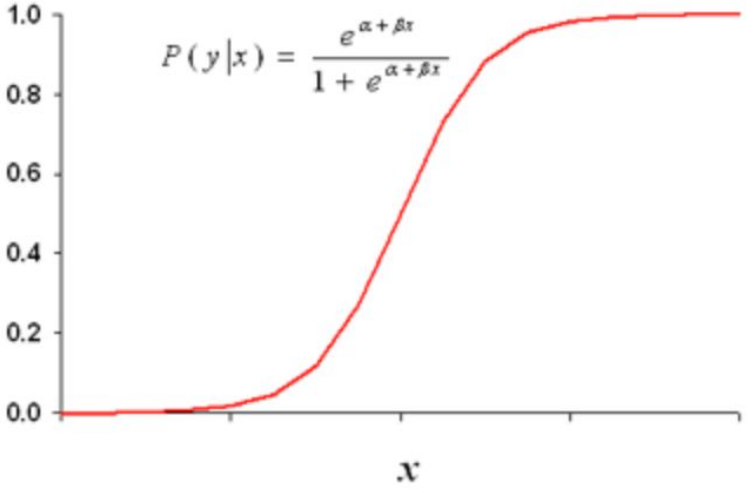
# 로짓(logit)변환과 오즈(odds)

- 다중선형 회귀분석 : y = b0 + b1x1 + b2x2+ … + bnxn , $-\infty < y < \infty$
- 로지스틱 회귀분석 : ln(p/1-p) = b0+ b1x1 + b2x2+ … + bnxn, $-\infty < \ln(p/1-p) < \infty$
  - 독립변수가 1개인 경우 로지스틱 회귀분석은

$$\ln\left(\frac{P}{1-P}\right) = a + bX$$

$$\frac{P}{1-P} = e^{a+bX}$$

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

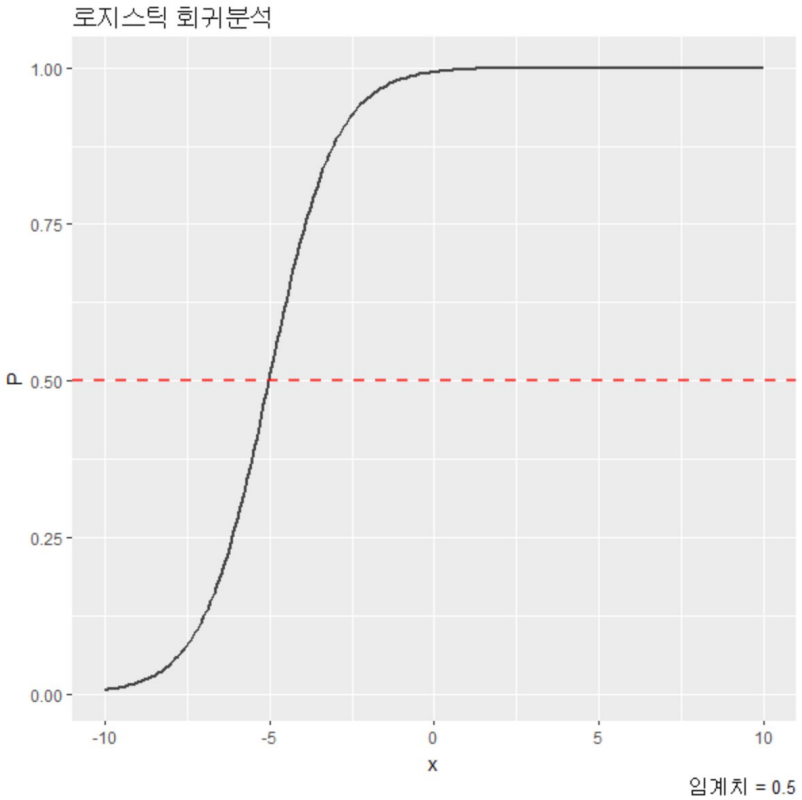


# 최대우도추정법(Maximum Likelihood Estimation, MLE)

- 선형회귀 모형에서는 최소제곱법을 사용했으나, 로지스틱에서 y는 0 또는 1을 갖기 때문에 다른 손실 함수가 필요
- 이 때 MLE를 사용하며, cross-entropy를 최소화 하는 과정과 동일함
- 참고자료
  - MLE : https://angeloyeo.github.io/2020/07/17/MLE.html
  - 로지스틱 회귀 파라메터 추정 : https://ratsgo.github.io/machine%20learning/2017/07/02/logistic/
  - 딥러닝 모델의 손실함수 : https://ratsgo.github.io/deep%20learning/2017/09/24/loss/

# 최적의 임계치

- 주로 p=0.5 를 기준으로 0.5보다  크면 A로 분류, 작으면 B로 분류
- 임계치가 0.5로 정해져 있는건 아니기 때문에 더 잘 분류할 수 있는 임계치를 찾아야 함
- 최적의 임계치는 Train Data를 기준으로 만든 회귀식에서 Validation Data를 넣어 더 잘 분류하는 임계치가 얼마인지를 확인해서 결정



# 로지스틱 회귀의 장점

- 재계산 없이 새 데이터에 대해 빨리 결과를 계산할 수 있음
- 모델의 해석이 다른 분류 방법들 보다 쉬움
  - 해석을 하기 위해서는 오즈비를 잘 이해해야함!