로지스틱 회귀

Logistic regression

다중선형회귀(Multiple Linear Regression)

- **다중선형회귀(Multiple Linear Regression)란?**
  - 수치형 설명변수(X) 와 연속형 숫자로 이뤄진 종속변수(Y) 간의 관계를 선형으로 가정하고 이를 가장 잘 표현할 수 있는 회귀 계수(β)를 데이터로부터 추정하는 것
  - 다중선형회귀모델 (오차항: ϵ, 회귀계수: β_p)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

- **회귀 계수 결정법**
  - 회귀 계수들은 모델의 예측 값($\bar{Y}$)과 실제 값(Y)의 차이, 즉 **오차 제곱 합(error sum of squares)**을 최소로 하는 값으로 결정 가능함
  - 즉, 회귀 계수에 대해 미분한 식을 0으로 놓고 풀어 **명시적인 해**를 구할 수 있음

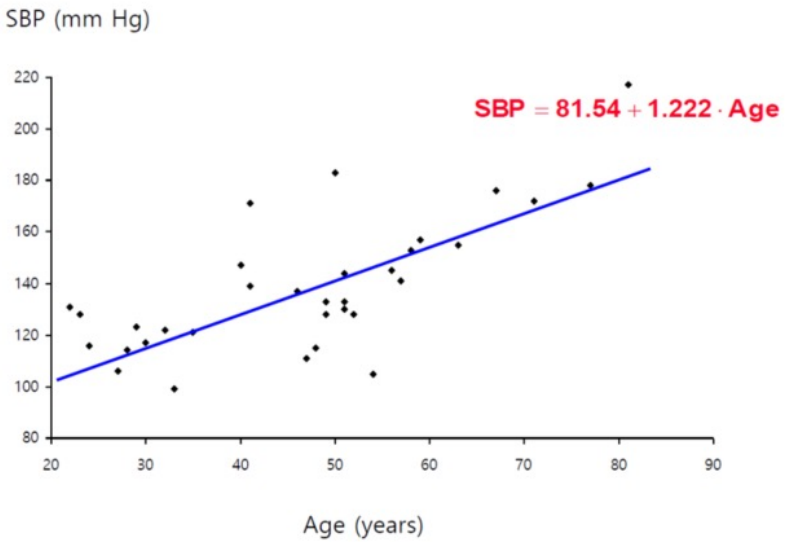
다중선형회귀(Multiple Linear Regression)

- **[다중선형회귀 예시 #1]**
  - 33명의 성인 여성에 대한 나이와 혈압 데이터
  - 오차제곱합을 최소로 하는 회귀 계수 계산 결과 및 분석
    - SBP = 81.54 + 1.222AGE
    - 나이라는 변수에 대응하는 계수는 1.222로 나타났는데, 이는 나이를 한 살 더 먹으면 혈압이 1.222mm/Hg만큼 증가한다는 결과를 보여줌

Age	SBP
22	131
23	128
24	116
27	106
28	114
29	123
30	117
32	122
33	99
35	121
40	147

Age	SBP
41	139
41	171
46	137
47	111
48	115
49	133
49	128
50	183
51	130
51	133
51	144

Age	SBP
52	128
54	105
56	145
57	141
58	153
59	157
63	155
67	176
71	172
77	178
81	217



다중선형회귀(Multiple Linear Regression)

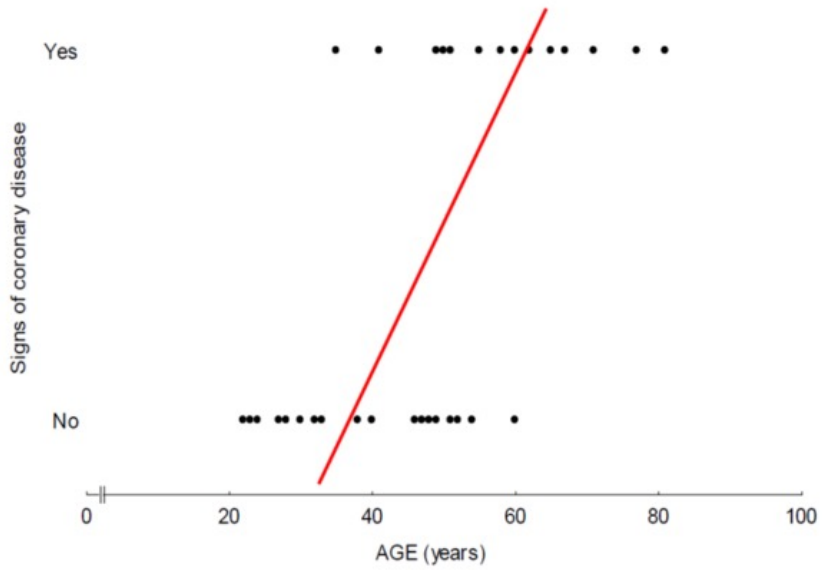
- **[다중선형회귀 예시 #2]**
  - 33명의 성인 여성에 대한 나이(Age)와 암 발병(CD) 데이터
  - 오차제곱합을 최소로 하는 회귀 계수 계산 결과 및 분석
  - 다중선형회귀 모델 적용 불가
    - ∵ 범주형 숫자(암 발병 여부)는 연속형 숫자(혈압)와 달리 의미를 지니지 않음
    - 즉, 0(정상)과 1(발병)을 서로 바꾸어도 상관 없음

➔ 범주형 숫자는 로지스틱 회귀 모델 적용 가능

Age	CD
22	0
23	0
24	0
27	0
28	0
30	0
30	0
32	0
33	0
35	1
38	0

Age	CD
40	0
41	1
46	0
47	0
48	0
49	1
49	0
50	1
51	0
51	1
52	0

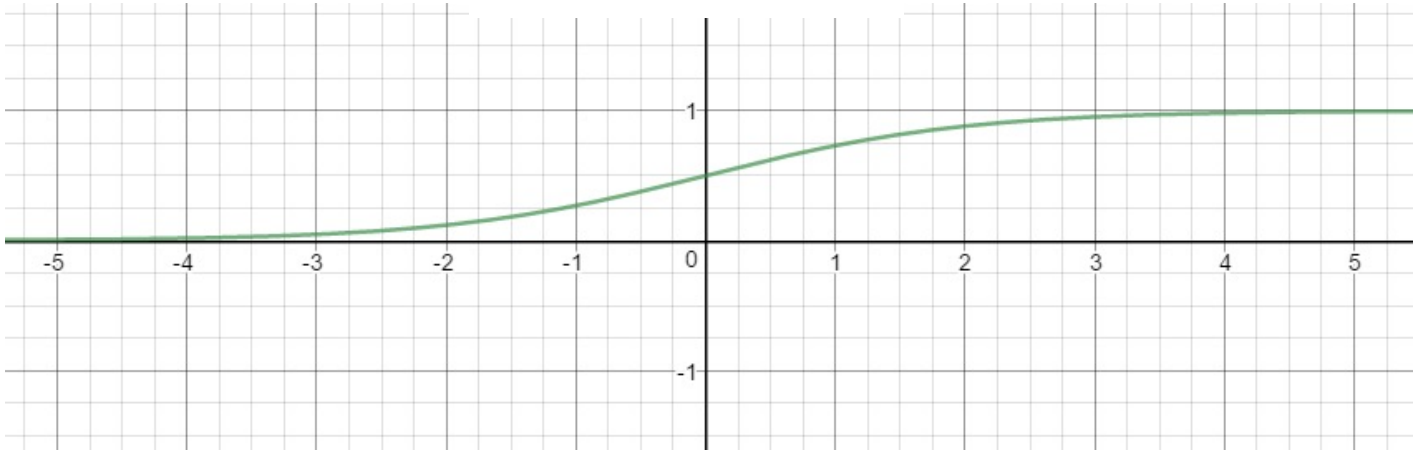
Age	CD
54	0
55	1
58	1
60	1
60	0
62	1
65	1
67	1
71	1
77	1
81	1



로지스틱 함수 (Logistic function)

- **로지스틱 함수 란?**
  - 아래 그림과 같이 S-커브 함수를 나타냄
    - 실제 많은 자연, 사회현상에서는 특정 변수에 대한 확률 값이 선형이 아닌 S-커브 형태를 따르는 경우가 많음
  - x값으로 어떤 값이든 받을 수가 있지만 출력 결과(y)는 항상 0에서 1사이 값이 됨
    - 누적분포함수(cumulative distribution function) 요건을 충족
  - 시그모이드 함수라고 명명하기도 함

$$y = \frac{1}{1 + e^{-x}}$$

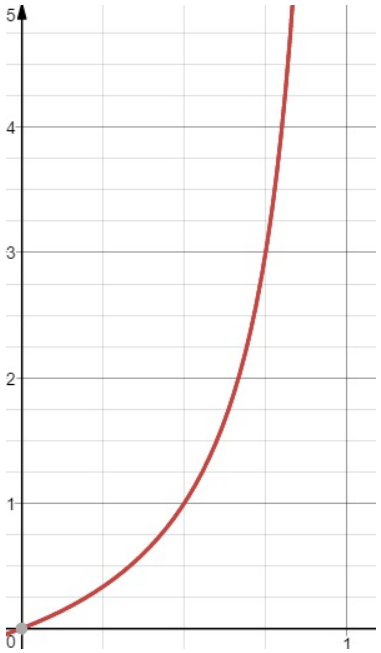


승산 (Odds)

- **승산 (Odds)이란?**
  - 임의의 사건 A가 발생하지 않을 확률 대비 일어날 확률의 비율

$$odds = \frac{P(A)}{P(A^c)} = \frac{P(A)}{1 - P(A)}$$

  - P(A)가 1에 가까울 수록 승산은 커지고, 반대로 P(A)가 0이라면 승산은 0



이항 로지스틱 회귀

- Y가 범주형 일 경우, 회귀 모델을 적용할 수 없음
- Y를 확률식으로 바꿔보면

$$P(Y = 1|X = \vec{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$
$$= \vec{\beta}^T \vec{x}$$

- Y를 승산으로 바꿔보면

$$\frac{P(Y = 1|X = \vec{x})}{1 - P(Y = 1|X = \vec{x})} = \vec{\beta}^T \vec{x}$$

- Y 승산에 로그를 취하면

$$\log\left(\frac{P(Y = 1|X = \vec{x})}{1 - P(Y = 1|X = \vec{x})}\right) = \vec{\beta}^T \vec{x}$$

이항 로지스틱 회귀

- x가 주어졌을 때 범주1일 확률을 p(x), 위 식 우변을 a로 치환해 정리하면

$$\frac{p(x)}{1-p(x)} = e^a$$

$$\begin{aligned} p(x) &= e^a \{1-p(x)\} \\ &= e^a - e^a p(x) \end{aligned}$$

$$p(x)\left(1+e^a\right) = e^a$$

$$p(x) = \frac{e^a}{1+e^a} = \frac{1}{1+e^{-a}}$$

$$\therefore P(Y=1|X=\vec{x}) = \frac{1}{1+e^{-\vec{\beta}^T\vec{x}}}$$

➔ 로지스틱 함수

이항 로지스틱 회귀의 결정 경계

- 이항 로지스틱 모델은 범주 정보를 모르는 입력 벡터 x를 넣으면 범주 1에 속할 확률을 반환하며, 범주 1로 분류하는 판단 기준은 아래와 같음

$$P(Y = 1|X = \vec{x}) > P(Y = 0|X = \vec{x})$$

- 범주가 두 개 뿐이므로, 위 식 좌변을 p(x)로 치환하면,

$$p(x) > 1 - p(x)$$

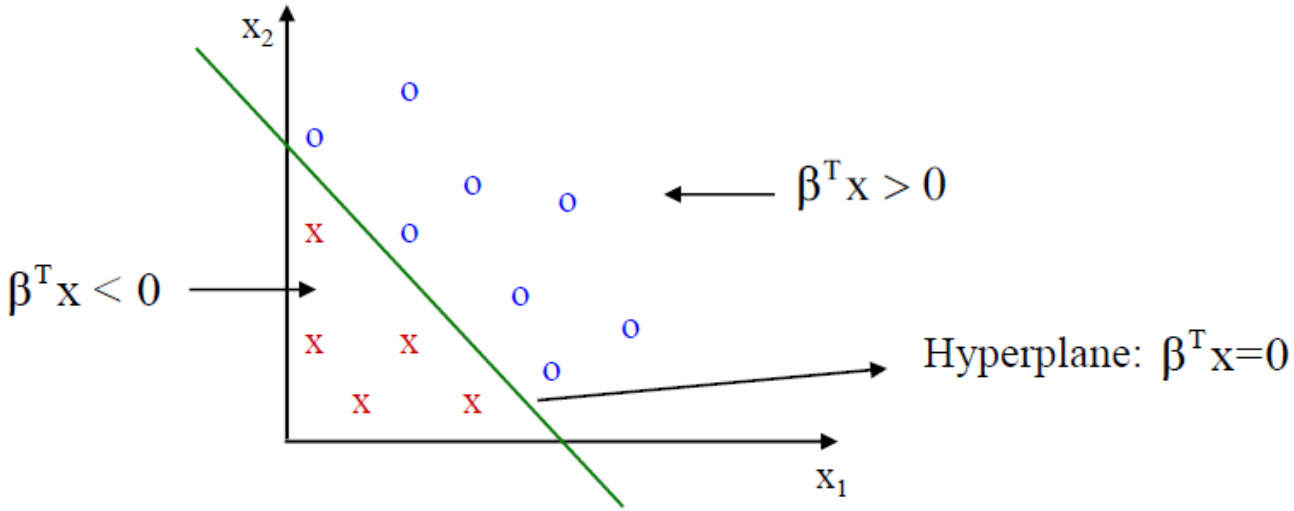
$$\frac{p(x)}{1 - p(x)} > 1$$

$$\log \frac{p(x)}{1 - p(x)} > 0$$

$$\therefore \vec{\beta}^T \vec{x} > 0$$

- 마찬가지로 $\beta^T x < 0$ 이면 데이터 범주를 0으로 분류하게 되며, 로지스틱 결정 경계 (decision boundary)는 $\beta^T x = 0$ 인 하이퍼플레인(hyperplane) 입니다.

이항 로지스틱 회귀의 결정 경계



Classifier

$$y = \frac{1}{\left(1+\exp(-\beta^T x)\right)}$$

$$\begin{pmatrix} y \to 1 & if & \beta^T x \to & \infty \\ y = \frac{1}{2} & if & \beta^T x = & 0 \\ y \to 0 & if & \beta^T x \to & -\infty \end{pmatrix}$$

다항 로지스틱 회귀

- 이항 로지스틱 회귀 모델을 통한 다항 로지스틱 회귀 문제 풀기

$$\log \frac{P(Y=1|X=\vec{x})}{P(Y=3|X=\vec{x})} = \beta_1^T \vec{x}$$

$$\log \frac{P(Y=2|X=\vec{x})}{P(Y=3|X=\vec{x})} = \beta_2^T \vec{x}$$

- 세번째 범주에 속할 확률 = 1 – 첫번째 범주에 속할 확률 – 두번째 범주에 속할 확률

$$P(Y=1|X=\vec{x}) = \frac{e^{\vec{\beta_1}^T \vec{x}}}{1 + e^{\vec{\beta_1}^T \vec{x}} + e^{\vec{\beta_2}^T \vec{x}}}$$

$$P(Y=2|X=\vec{x}) = \frac{e^{\vec{\beta_2}^T \vec{x}}}{1 + e^{\vec{\beta_1}^T \vec{x}} + e^{\vec{\beta_2}^T \vec{x}}}$$

$$P(Y=3|X=\vec{x}) = \frac{1}{1 + e^{\vec{\beta_1}^T \vec{x}} + e^{\vec{\beta_2}^T \vec{x}}}$$

- K개 범주를 분류하는 다항로지스틱 회귀 모델의 입력 벡터 x가 각 클래스로 분류될 확률

$$P(Y=k|X=\vec{x}) = \frac{e^{\vec{\beta_k}^T \vec{x}}}{1 + \sum_{i=1}^{K-1} e^{\vec{\beta_i}^T \vec{x}}} \quad (k = 0, 1, \ldots, K-1)$$

$$P(Y=K|X=\vec{x}) = \frac{1}{1 + \sum_{i=1}^{K-1} e^{\vec{\beta_i}^T \vec{x}}}$$

다항 로지스틱 회귀와 소프트맥스

- '로그승산'으로 된 좌변을 '로그확률'로 변경

$$P(Y=1|X=\vec{x}) = \frac{e^{\vec{\beta_1}^T \vec{x}}}{1 + e^{\vec{\beta_1}^T \vec{x}} + e^{\vec{\beta_2}^T \vec{x}}}$$

$$P(Y=2|X=\vec{x}) = \frac{e^{\vec{\beta_2}^T \vec{x}}}{1 + e^{\vec{\beta_1}^T \vec{x}} + e^{\vec{\beta_2}^T \vec{x}}}$$

$$P(Y=3|X=\vec{x}) = \frac{1}{1 + e^{\vec{\beta_1}^T \vec{x}} + e^{\vec{\beta_2}^T \vec{x}}}$$

$$\log P(Y=1|X=\vec{x}) = \vec{\beta_1}^T \vec{x} - \log Z$$

$$\log P(Y=2|X=\vec{x}) = \vec{\beta_2}^T \vec{x} - \log Z$$

$$\cdots$$

$$\log P(Y=K|X=\vec{x}) = \vec{\beta_K}^T \vec{x} - \log Z$$

- 로그 성질을 활용해 c번째 범주에 속할 확률을 기준으로 식을 정리

$$\log P(Y=c) + \log Z = \vec{\beta_c}^T \vec{x}$$

$$\log \{P(Y=c) \times Z\} = \vec{\beta_c}^T \vec{x}$$

$$P(Y=c) \times Z = e^{\vec{\beta_c}^T \vec{x}}$$

$$P(Y=c) = \frac{1}{Z} e^{\vec{\beta_c}^T \vec{x}}$$

$$P(Y=c) = \frac{e^{\vec{\beta_c}^T \vec{x}}}{\sum_{k=1}^{K} e^{\vec{\beta_k}^T \vec{x}}}$$

- 전체 확률 합은 1

$$1 = \sum_{k=1}^{K} P(Y=k) = \sum_{k=1}^{K} \frac{1}{Z} e^{\vec{\beta_k}^T \vec{x}} = \frac{1}{Z} \sum_{k=1}^{K} e^{\vec{\beta_k}^T \vec{x}} \quad \therefore Z = \sum_{k=1}^{K} e^{\vec{\beta_k}^T \vec{x}}$$