

# 로지스틱 회귀분석 (Logistic Regression)

SCH Univ.  
Dept. of AI and Bigdata  
Lee Howoo

# 개요

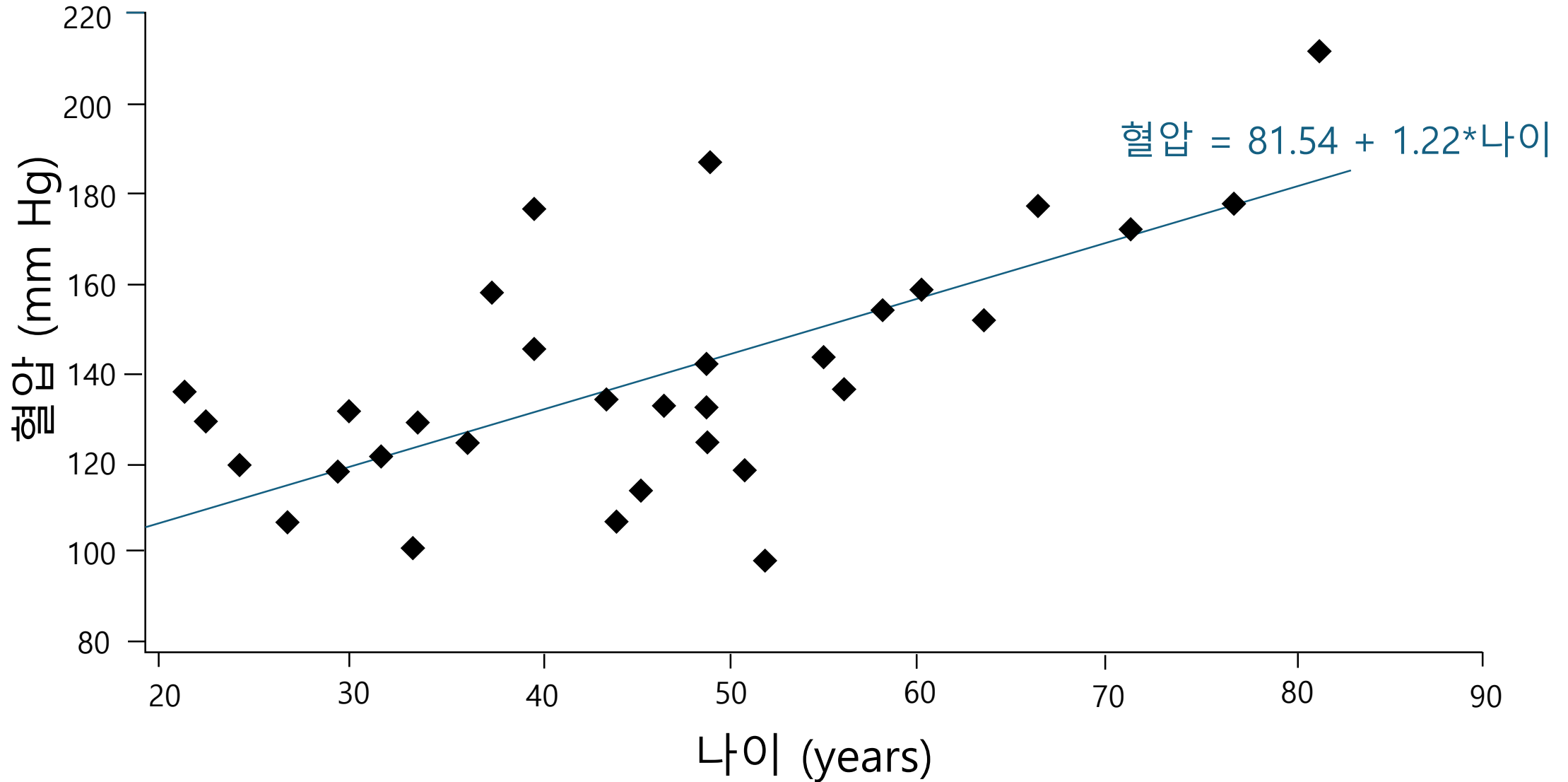
- 로지스틱 회귀모델
- 로지스틱 회귀모델 형태
- 아드 (Odds)
- 파라미터 추정
- 로지스틱 회귀모델 결과 및 해석
- 로지스틱 회귀모델 예제

# 로지스틱 회귀모델 배경

- 33명의 나이와 혈압의 관계

나이	혈압	나이	혈압	나이	혈압
22	131	41	139	52	128
23	128	41	171	54	105
24	116	46	137	56	145
27	106	47	111	57	141
28	114	48	115	58	153
29	123	49	133	59	157
30	117	49	128	63	155
32	122	50	183	67	176
33	99	51	130	71	172
35	121	51	133	77	178
40	147	51	144	81	217

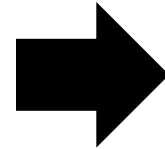
# 로지스틱 회귀모델 배경



# 로지스틱 회귀분석

온도	판매량
20	40
21	42
22	44
23	46

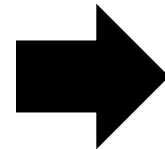
양적



회귀  
regression

공부시간	시험결과
20	불합격
21	불합격
22	합격
23	합격

범주형

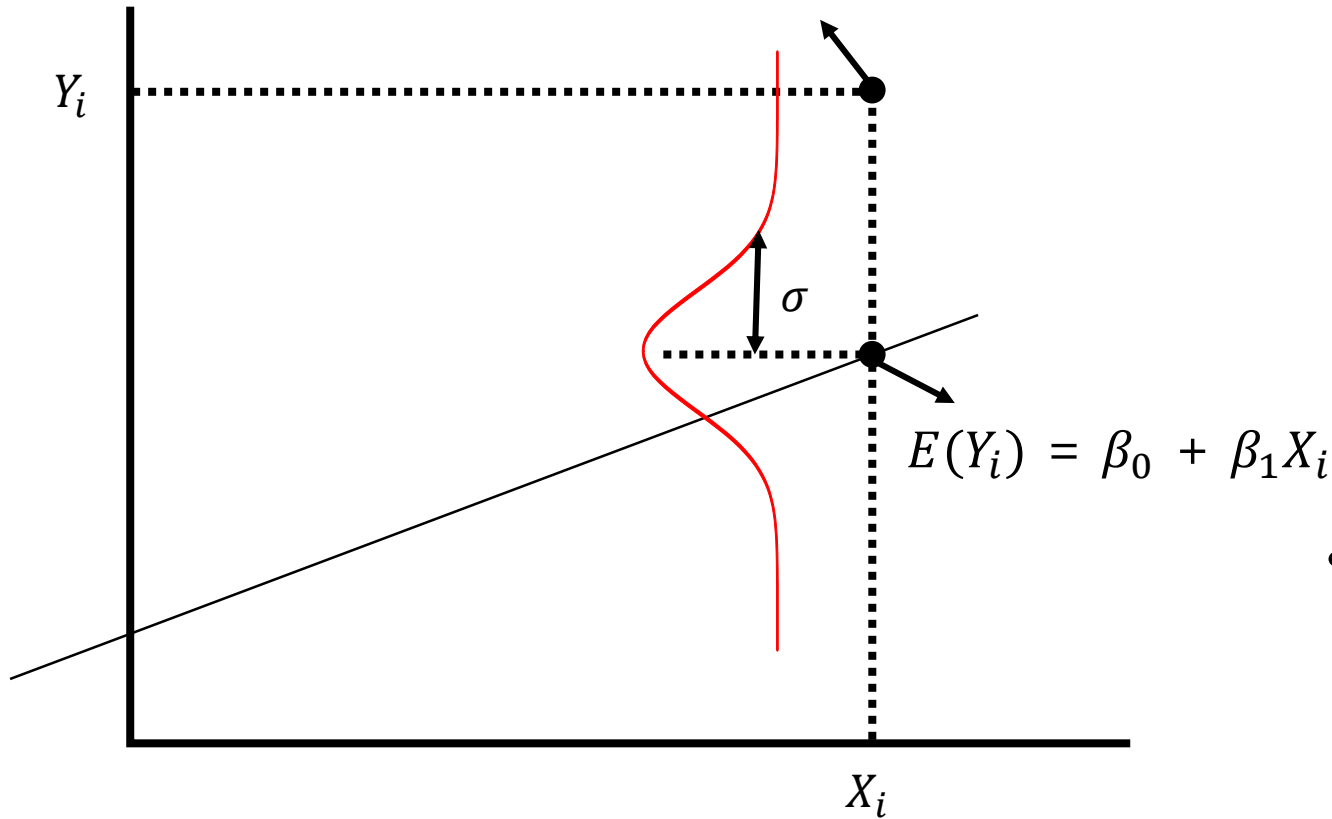


분류  
classification

로지스틱 회귀분석은 이름에는 회귀가 들어갔지만, **분류 모델임**

# 로지스틱 회귀모델과 선형회귀 모델의 차이

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$



- 선형 회귀의 대전제  
“확률오차가 정규분포를 따름”
  - 위 전제를 바탕으로  
“Y값도 정규분포를 따름”
  - 위 전제를 바탕으로 파라미터 값도  
최소제곱법으로 추정
- 선형 회귀와 로지스틱의 차이점
  - 가장 큰 차이점 → Y값이 0 또는 1인  
값, 즉 Y값이 정규분포가 아님
  - 로지스틱에서는 범주형 Y값은 정규분  
포가 아니라 베르누이 분포 형태를 띠

# 베르누이 분포

베르누이 시행: 결과가 두 가지 중 하나로만 나오는 실험이나 시행  
ex) 동전 던지기

베르누이 확률변수: 베르누이 시행 결과를 실수 0 또는 1로 바꾼 것

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

$$\text{Bern}(x; \pi) = \pi^x (1 - \pi)^{(1-x)}$$

베르누이 분포에서 1이 나올 확률을 의미하는  $\pi$ 는 모수(parameter)를 가짐  
변수와 모수는 세미콜론 기호로 분리해서 표기

# 로지스틱 회귀모델 이론 배경

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad Y_i = 0 \text{ or } 1$$



# 로지스틱 회귀모델 이론 배경

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad Y_i = 0 \text{ or } 1$$

Assume  $E(\varepsilon_i) = 0$

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

# 로지스틱 회귀모델 이론 배경

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad Y_i = 0 \text{ or } 1$$

$$\text{Assume } E(\varepsilon_i) = 0 \quad E(Y_i) = \beta_0 + \beta_1 X_i$$

Consider  $Y_i$  to be a Bernulli random variable

$$P(Y_i = 1) = \pi_i$$

$$P(Y_i = 0) = 1 - \pi_i$$

# 로지스틱 회귀모델 이론 배경

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad Y_i = 0 \text{ or } 1$$

$$\text{Assume } E(\varepsilon_i) = 0 \quad E(Y_i) = \beta_0 + \beta_1 X_i$$

Consider  $Y_i$  to be a Bernulli random variable

$$P(Y_i = 1) = \pi_i$$

$$P(Y_i = 0) = 1 - \pi_i$$

$$E(Y_i) = 1 \cdot \pi_i + 0 \cdot (1 - \pi_i) = \pi_i$$

# 로지스틱 회귀모델 이론 배경

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad Y_i = 0 \text{ or } 1$$

$$\text{Assume } E(\varepsilon_i) = 0 \quad E(Y_i) = \beta_0 + \beta_1 X_i$$

Consider  $Y_i$  to be a Bernulli random variable

$$P(Y_i = 1) = \pi_i$$

$$P(Y_i = 0) = 1 - \pi_i$$

$$E(Y_i) = 1 \cdot \pi_i + 0 \cdot (1 - \pi_i) = \pi_i$$

$$E(Y_i) = \beta_0 + \beta_1 X_i = \pi_i$$

X값이 주어졌을 때 출력변수 Y가 1의 값을 가질 확률

# 로지스틱 회귀모델

- 출력변수가 연속형이 아닌 이진 범주형 질병유무?

나이	질병유무	나이	질병유무	나이	질병유무
22	0	41	0	52	0
23	0	41	1	54	1
24	0	46	0	56	1
27	0	47	0	57	1
28	0	48	0	58	0
29	0	49	1	59	1
30	0	49	0	63	1
32	0	50	1	67	1
33	0	51	0	71	1
35	1	51	1	77	1
40	0	51	0	81	1

# 로지스틱 회귀모델

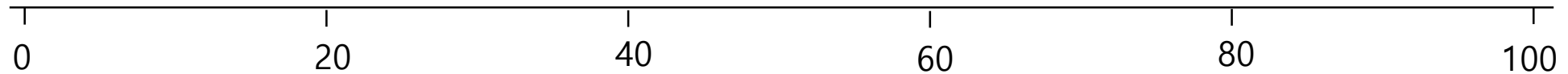
- 두 변수 사이의 관계식은 선형?

Yes (1) -



무  
양  
변  
질

No (0) -



나이 (years)

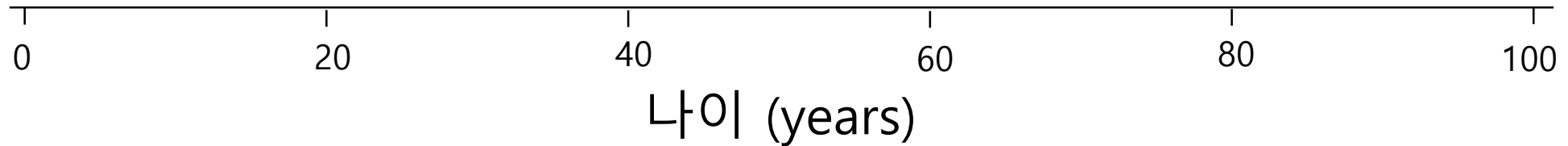
# 로지스틱 회귀모델

- 두 변수 사이의 관계식은 선형?

Yes (1) -

무  
양  
분  
리

No (0) -

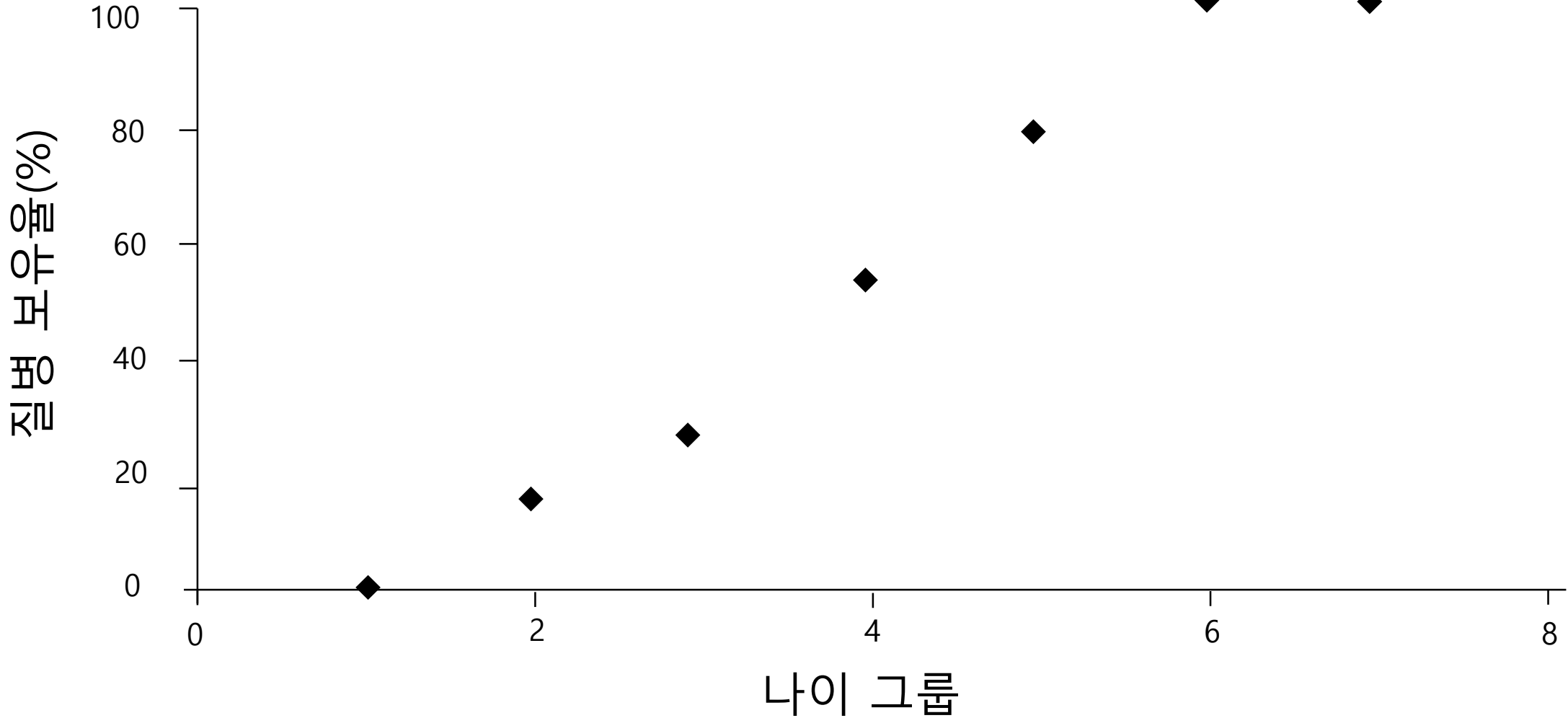


# 로지스틱 회귀모델

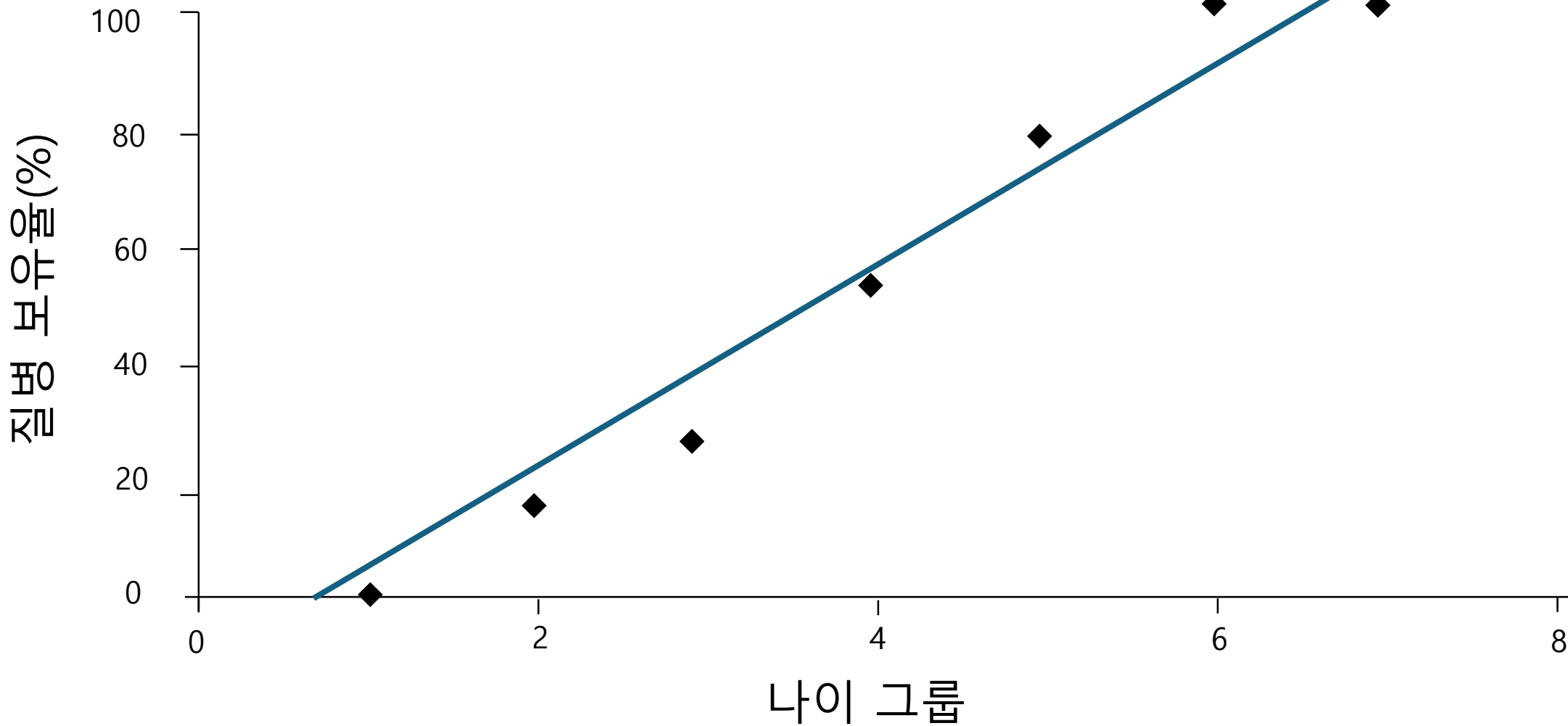
나이 그룹	그룹내 수	질병	
		질병보유자 수	%
20 - 29	5	0	0
30 - 39	6	1	17
40 - 49	7	2	29
50 - 59	7	4	57
60 - 69	5	4	80
70 - 79	2	2	100
80 - 89	1	1	100



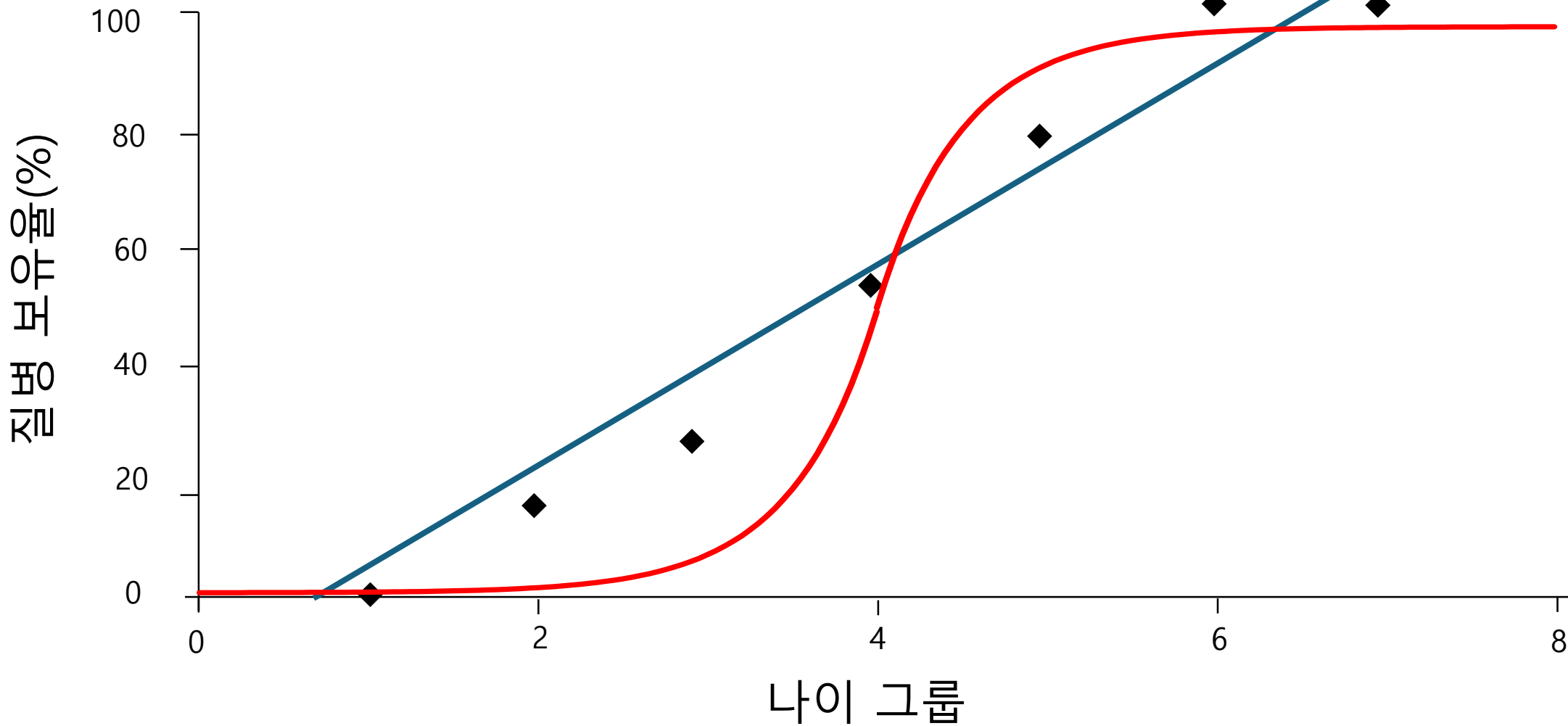
# 로지스틱 회귀분석 알고리즘



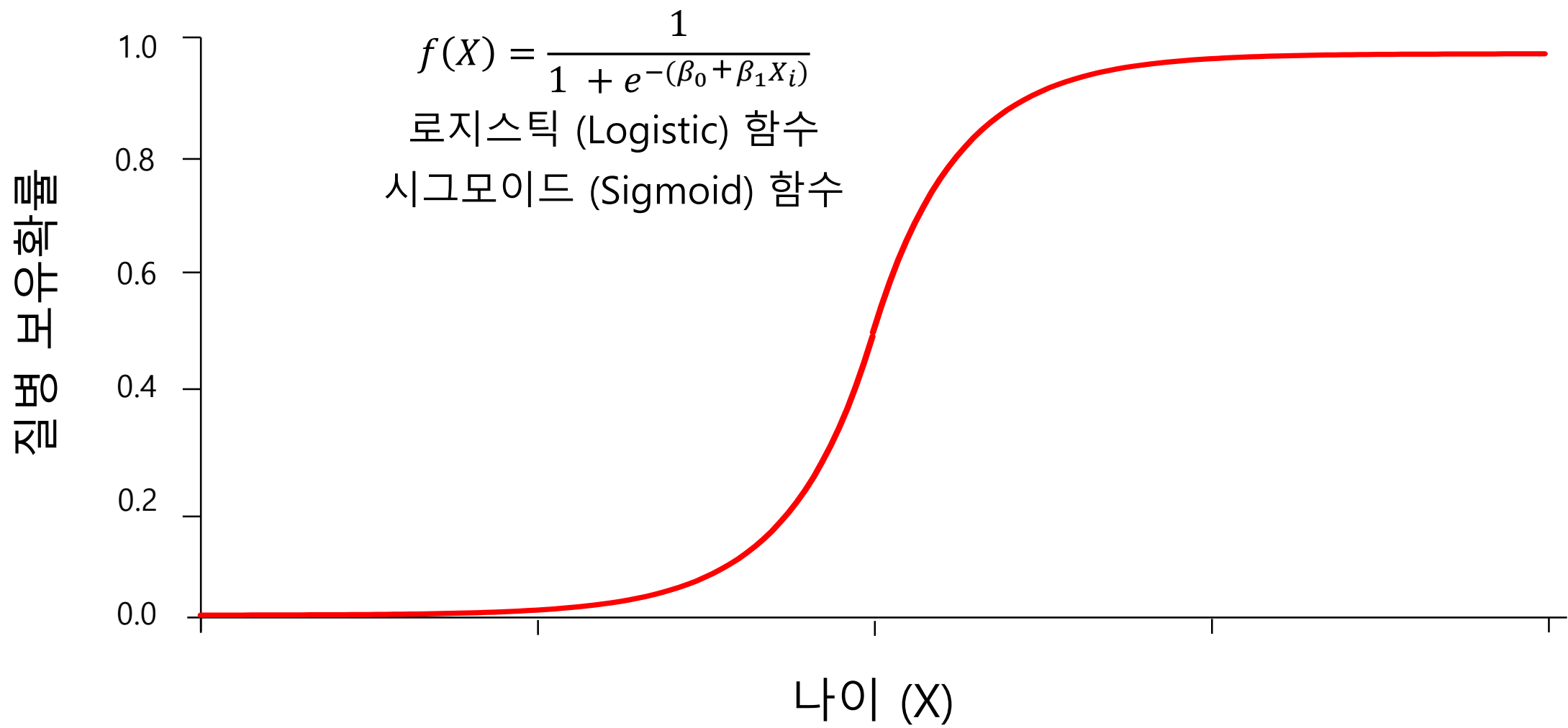
# 로지스틱 회귀분석 알고리즘



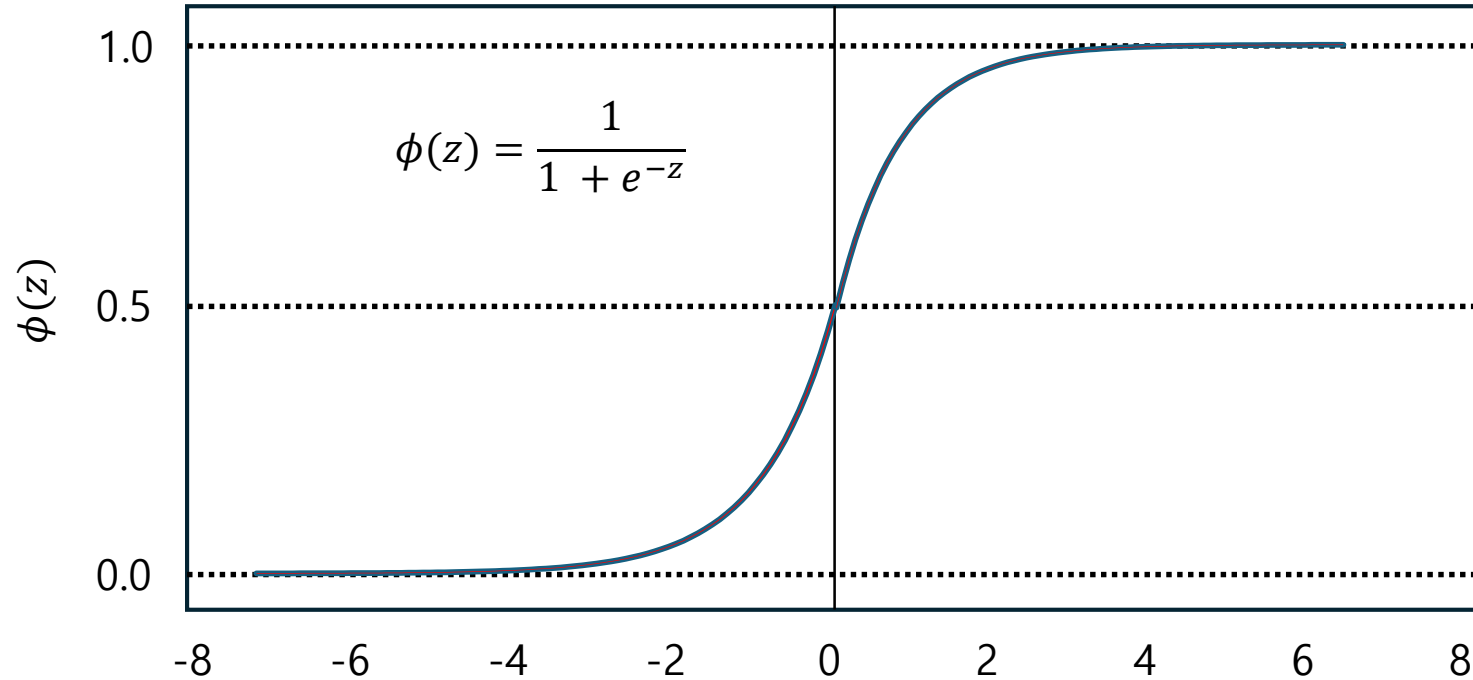
# 로지스틱 회귀분석 알고리즘



# 로지스틱 회귀분석 알고리즘 - 로지스틱 함수



# 로지스틱 회귀분석 알고리즘 - 로지스틱 함수



- Logistic function, Sigmoid function, Squashing function (Large input  $\rightarrow$  Small output)
- 아웃풋 범위 : 0 ~ 1
- 인풋값에 대해 단조증가 (혹은 단조감소) 함수
- 미분결과를 아웃풋의 함수로 표현 가능 (Gradient learning method에 유용하게 사용)
- $\frac{d\phi(z)}{dx} = \frac{1}{1 + e^{-z}} \left( 1 - \frac{1}{1 + e^{-z}} \right) = \phi(z)(1 - \phi(z))$

# 로지스틱 회귀분석 알고리즘 – 로지스틱 함수

$$E(y) = \pi(X = x) = P(Y = 1|X = x) = 1 - P(Y = 0|X = x)$$

단순 로지스틱 회귀모델: 입력변수  $X$ 가 1개인 로지스틱 회귀모델

$$E(y) = \pi(X = x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_i)}}$$

관측치  $x$ 가 범주 1에 속할 확률

(Probability that an observation  $x$  belongs to class 1)

# 로지스틱 회귀모델 - $\beta_1$ 의 해석

$$E(y) = \pi(X = x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

$\beta_1$ 의 해석  $\rightarrow$  직관적이지 못함

승산 (Odds)

- 성공확률을  $p$ 로 정의할 때, 실패 대비 성공 확률 비율

$$\text{Odd} = \frac{p}{1-p}$$

$$p = 1 \rightarrow \text{odd} = \infty$$

$$p = 0 \rightarrow \text{odd} = 0$$

# 로지스틱 회귀모델 - $\beta_1$ 의 해석 - Odds



FIFA WORLD CUP  
RUSSIA 2018

2018 FIFA WORLD CUP					
June 14th – July 15th, 2018					
Various Locations throughout Russia					
ODDS TO WIN:	OPENING ODDS 7/14/2014	CURRENT ODDS 2/12/2018	ODDS TO WIN:	OPENING ODDS 7/14/2014	CURRENT ODDS 2/12/2018
<b>96101 GERMANY</b>	<b>5/1</b>	<b>4/1</b>	<b>96117 SWEDEN</b>	<b>80/1</b>	<b>100/1</b>
<b>96102 BRAZIL</b>	<b>8/1</b>	<b>5/1</b>	<b>96118 SERBIA</b>	<b>100/1</b>	<b>150/1</b>
<b>96103 FRANCE</b>	<b>10/1</b>	<b>11/2</b>	<b>96119 SENEGAL</b>	<b>500/1</b>	<b>150/1</b>
<b>96104 SPAIN</b>	<b>8/1</b>	<b>13/2</b>	<b>96120 EGYPT</b>	<b>500/1</b>	<b>200/1</b>
<b>96105 ARGENTINA</b>	<b>8/1</b>	<b>7/1</b>	<b>96121 ICELAND</b>	<b>1000/1</b>	<b>200/1</b>
<b>96106 BELGIUM</b>	<b>15/1</b>	<b>10/1</b>	<b>96122 PERU</b>	<b>500/1</b>	<b>200/1</b>
<b>96107 ENGLAND</b>	<b>25/1</b>	<b>15/1</b>	<b>96123 NIGERIA</b>	<b>150/1</b>	<b>200/1</b>
<b>96108 PORTUGAL</b>	<b>30/1</b>	<b>20/1</b>	<b>96124 JAPAN</b>	<b>150/1</b>	<b>300/1</b>
<b>96109 URUGUAY</b>	<b>50/1</b>	<b>30/1</b>	<b>96125 COSTA RICA</b>	<b>200/1</b>	<b>250/1</b>
<b>96110 COLOMBIA</b>	<b>20/1</b>	<b>40/1</b>	<b>96126 AUSTRALIA</b>	<b>300/1</b>	<b>300/1</b>
<b>96111 RUSSIA</b>	<b>20/1</b>	<b>40/1</b>	<b>96127 MOROCCO</b>	<b>500/1</b>	<b>500/1</b>
<b>96112 CROATIA</b>	<b>60/1</b>	<b>40/1</b>	<b>96128 IRAN</b>	<b>2000/1</b>	<b>500/1</b>
<b>96113 POLAND</b>	<b>100/1</b>	<b>40/1</b>	<b>96129 SOUTH KOREA</b>	<b>200/1</b>	<b>500/1</b>
<b>96114 MEXICO</b>	<b>50/1</b>	<b>60/1</b>	<b>96130 TUNISIA</b>	<b>500/1</b>	<b>1000/1</b>
<b>96115 DENMARK</b>	<b>100/1</b>	<b>100/1</b>	<b>96131 PANAMA</b>	<b>1000/1</b>	<b>1000/1</b>
<b>96116 SWITZERLAND</b>	<b>80/1</b>	<b>100/1</b>	<b>96132 SAUDI ARABIA</b>	<b>1000/1</b>	<b>1000/1</b>



# 로지스틱 회귀모델 - $\beta_1$ 의 해석 - Odds



FIFA WORLD CUP  
RUSSIA 2018

2018 FIFA WORLD CUP					
June 14th – July 15th, 2018					
Various Locations throughout Russia					
ODDS TO WIN:	OPENING ODDS 7/14/2014	CURRENT ODDS 2/12/2018	ODDS TO WIN:	OPENING ODDS 7/14/2014	CURRENT ODDS 2/12/2018
<b>96101 GERMANY</b>	<b>5/1</b>	<b>4/1</b>	<b>96117 SWEDEN</b>	<b>80/1</b>	<b>100/1</b>
<b>96102 BRAZIL</b>	<b>8/1</b>	<b>5/1</b>	<b>96118 SERBIA</b>	<b>100/1</b>	<b>150/1</b>
<b>96103 FRANCE</b>	<b>10/1</b>	<b>11/2</b>	<b>96119 SENEGAL</b>	<b>500/1</b>	<b>150/1</b>
<b>96104 SPAIN</b>	<b>8/1</b>	<b>13/2</b>	<b>96120 EGYPT</b>	<b>500/1</b>	<b>200/1</b>
<b>96105 ARGENTINA</b>	<b>8/1</b>	<b>7/1</b>	<b>96121 ICELAND</b>	<b>1000/1</b>	<b>200/1</b>
<b>96106 BELGIUM</b>	<b>15/1</b>	<b>10/1</b>	<b>96122 PERU</b>	<b>500/1</b>	<b>200/1</b>
<b>96107 ENGLAND</b>	<b>25/1</b>	<b>15/1</b>	<b>96123 NIGERIA</b>	<b>150/1</b>	<b>200/1</b>
<b>96108 PORTUGAL</b>	<b>30/1</b>	<b>20/1</b>	<b>96124 JAPAN</b>	<b>150/1</b>	<b>300/1</b>
<b>96109 URUGUAY</b>	<b>50/1</b>	<b>30/1</b>	<b>96125 COSTA RICA</b>	<b>200/1</b>	<b>250/1</b>
<b>96110 COLOMBIA</b>	<b>20/1</b>	<b>40/1</b>	<b>96126 AUSTRALIA</b>	<b>300/1</b>	<b>300/1</b>
<b>96111 RUSSIA</b>	<b>20/1</b>	<b>40/1</b>	<b>96127 MOROCCO</b>	<b>500/1</b>	<b>500/1</b>
<b>96112 CROATIA</b>	<b>60/1</b>	<b>40/1</b>	<b>96128 IRAN</b>	<b>2000/1</b>	<b>500/1</b>
<b>96113 POLAND</b>	<b>100/1</b>	<b>40/1</b>	<b>96129 SOUTH KOREA</b>	<b>200/1</b>	<b>500/1</b>
<b>96114 MEXICO</b>	<b>50/1</b>	<b>60/1</b>	<b>96130 TUNISIA</b>	<b>500/1</b>	<b>1000/1</b>
<b>96115 DENMARK</b>	<b>100/1</b>	<b>100/1</b>	<b>96131 PANAMA</b>	<b>1000/1</b>	<b>1000/1</b>
<b>96116 SWITZERLAND</b>	<b>80/1</b>	<b>100/1</b>	<b>96132 SAUDI ARABIA</b>	<b>1000/1</b>	<b>1000/1</b>

# 로지스틱 회귀모델 - $\beta_1$ 의 해석 - Odds



FIFA WORLD CUP  
RUSSIA 2018

2018 FIFA WORLD CUP					
June 14th – July 15th, 2018					
Various Locations throughout Russia					
ODDS TO WIN:	OPENING ODDS 7/14/2014	CURRENT ODDS 2/12/2018	ODDS TO WIN:	OPENING ODDS 7/14/2014	CURRENT ODDS 2/12/2018
<b>96101 GERMANY</b>	<b>5/1</b>	<b>4/1</b>	<b>96117 SWEDEN</b>	<b>80/1</b>	<b>100/1</b>
<b>96102 BRAZIL</b>	<b>8/1</b>	<b>5/1</b>	<b>96118 SERBIA</b>	<b>100/1</b>	<b>150/1</b>
<b>96103 FRANCE</b>	<b>10/1</b>	<b>11/2</b>	<b>96119 SENEGAL</b>	<b>500/1</b>	<b>150/1</b>
<b>96104 SPAIN</b>	<b>8/1</b>	<b>13/2</b>	<b>96120 EGYPT</b>	<b>500/1</b>	<b>200/1</b>
<b>96105 ARGENTINA</b>	<b>8/1</b>	<b>7/1</b>	<b>96121 ICELAND</b>	<b>1000/1</b>	<b>200/1</b>
<b>96106 BELGIUM</b>	<b>15/1</b>	<b>10/1</b>	<b>96122 PERU</b>	<b>500/1</b>	<b>200/1</b>
<b>96107 ENGLAND</b>	<b>25/1</b>	<b>15/1</b>	<b>96123 NIGERIA</b>	<b>150/1</b>	<b>200/1</b>
<b>96108 PORTUGAL</b>	<b>30/1</b>	<b>20/1</b>	<b>96124 JAPAN</b>	<b>150/1</b>	<b>300/1</b>
<b>96109 URUGUAY</b>	<b>50/1</b>	<b>30/1</b>	<b>96125 COSTA RICA</b>	<b>200/1</b>	<b>250/1</b>
<b>96110 COLOMBIA</b>	<b>20/1</b>	<b>40/1</b>	<b>96126 AUSTRALIA</b>	<b>300/1</b>	<b>300/1</b>
<b>96111 RUSSIA</b>	<b>20/1</b>	<b>40/1</b>	<b>96127 MOROCCO</b>	<b>500/1</b>	<b>500/1</b>
<b>96112 CROATIA</b>	<b>60/1</b>	<b>40/1</b>	<b>96128 IRAN</b>	<b>2000/1</b>	<b>500/1</b>
<b>96113 POLAND</b>	<b>100/1</b>	<b>40/1</b>	<b>96129 SOUTH KOREA</b>	<b>200/1</b>	<b>500/1</b>
<b>96114 MEXICO</b>	<b>50/1</b>	<b>60/1</b>	<b>96130 TUNISIA</b>	<b>500/1</b>	<b>1000/1</b>
<b>96115 DENMARK</b>	<b>100/1</b>	<b>100/1</b>	<b>96131 PANAMA</b>	<b>1000/1</b>	<b>1000/1</b>
<b>96116 SWITZERLAND</b>	<b>80/1</b>	<b>100/1</b>	<b>96132 SAUDI ARABIA</b>	<b>1000/1</b>	<b>1000/1</b>

# 로지스틱 회귀모델 - $\beta_1$ 의 해석 - Odds



프랑스의 우승 odds는 2/11

프랑스의 우승 확률은  $2/13 = 0.15(15\%)$



대한민국의 우승 odds는 1/500

대한민국의 우승 확률은  $1/501 \approx 0.001996 (0.1996\%)$

# 로지스틱 회귀모델 - $\beta_1$ 의 해석 - Odds



프랑스의 우승 odds는 2/11

프랑스의 우승 확률은  $2/13 = 0.15(15\%)$

$$\frac{P}{1-P} = \frac{2}{11}$$



대한민국의 우승 odds는 1/500

대한민국의 우승 확률은  $1/501 \approx 0.001996 (0.1996\%)$

# 로지스틱 회귀모델 - $\beta_1$ 의 해석 - Odds



프랑스의 우승 odds는 2/11

프랑스의 우승 확률은  $2/13 = 0.15(15\%)$

$$\frac{P}{1-P} = \frac{2}{11}$$
$$\rightarrow P = \frac{2}{13} = 0.15 (15\%)$$



대한민국의 우승 odds는 1/500

대한민국의 우승 확률은  $1/501 \approx 0.001996 (0.1996\%)$

# 로지스틱 회귀모델 - $\beta_1$ 의 해석 - Odds

$$\pi(X = x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \quad 0 \leq \pi(X = x) \leq 1$$

$$\text{Odds} = \frac{\pi(X=x)}{1 - \pi(X=x)}$$

Odds : 범주 0에 속할 확률 대비 범주 1에 속할 확률

(The ratio of the probability of belonging to class 1 to the probability of belonging to class 0)

# 로지스틱 회귀모델 - $\beta_1$ 의 해석 - Odds

$$\pi(X = x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \quad 0 \leq \pi(X = x) \leq 1$$

$$\text{Odds} = \frac{\pi(X=x)}{1 - \pi(X=x)}$$

Odds : 범주 0에 속할 확률 대비 범주 1에 속할 확률

(The ratio of the probability of belonging to class 1 to the probability of belonging to class 0)

$$\log(\text{Odds}) = \log\left(\frac{\pi(X = x)}{1 - \pi(X = x)}\right) :$$

# 로지스틱 회귀모델 - $\beta_1$ 의 해석 - Odds

$$\pi(X = x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \quad 0 \leq \pi(X = x) \leq 1$$

$$\text{Odds} = \frac{\pi(X=x)}{1 - \pi(X=x)}$$

Odds : 범주 0에 속할 확률 대비 범주 1에 속할 확률

(The ratio of the probability of belonging to class 1 to the probability of belonging to class 0)

$$\log(\text{Odds}) = \log\left(\frac{\pi(X = x)}{1 - \pi(X = x)}\right) = \log\left(\frac{\frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}}{1 + \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}}\right)$$



# 로지스틱 회귀모델 - $\beta_1$ 의 해석 - Odds

$$\pi(X = x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \quad 0 \leq \pi(X = x) \leq 1$$

$$\text{Odds} = \frac{\pi(X=x)}{1 - \pi(X=x)}$$

Odds : 범주 0에 속할 확률 대비 범주 1에 속할 확률

(The ratio of the probability of belonging to class 1 to the probability of belonging to class 0)

$$\log(\text{Odds}) = \log\left(\frac{\pi(X = x)}{1 - \pi(X = x)}\right) = \log\left(\frac{\frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}}{1 + \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}}\right) = \beta_0 + \beta_1 x$$

# 로지스틱 회귀모델 - $\beta_1$ 의 해석 - Odds

$$\pi(X = x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \quad 0 \leq \pi(X = x) \leq 1$$

$$\text{Odds} = \frac{\pi(X=x)}{1 - \pi(X=x)}$$

Odds : 범주 0에 속할 확률 대비 범주 1에 속할 확률

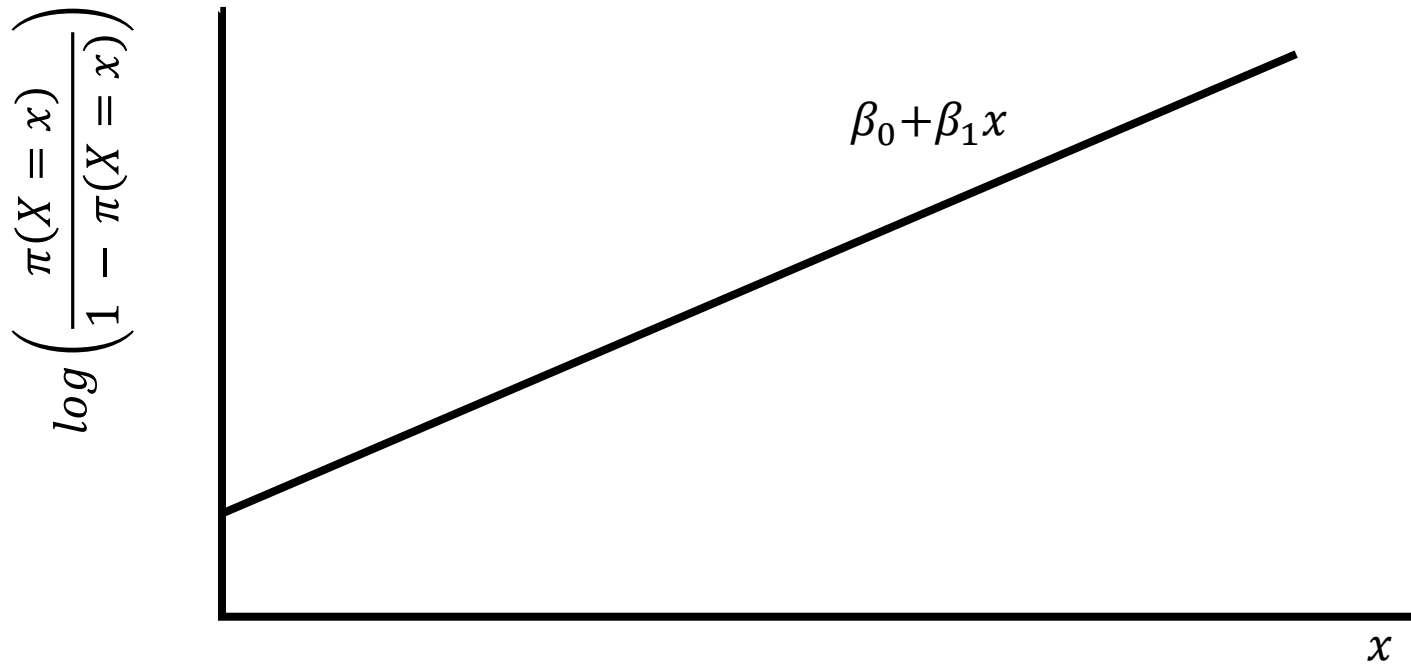
(The ratio of the probability of belonging to class 1 to the probability of belonging to class 0)

$$\log(\text{Odds}) = \log\left(\frac{\pi(X = x)}{1 - \pi(X = x)}\right) = \log\left(\frac{\frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}}{1 + \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}}\right) = \beta_0 + \beta_1 x$$

Logit Transform (로짓 변환)

# 로지스틱 회귀모델 - $\beta_1$ 의 해석 - Odds

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



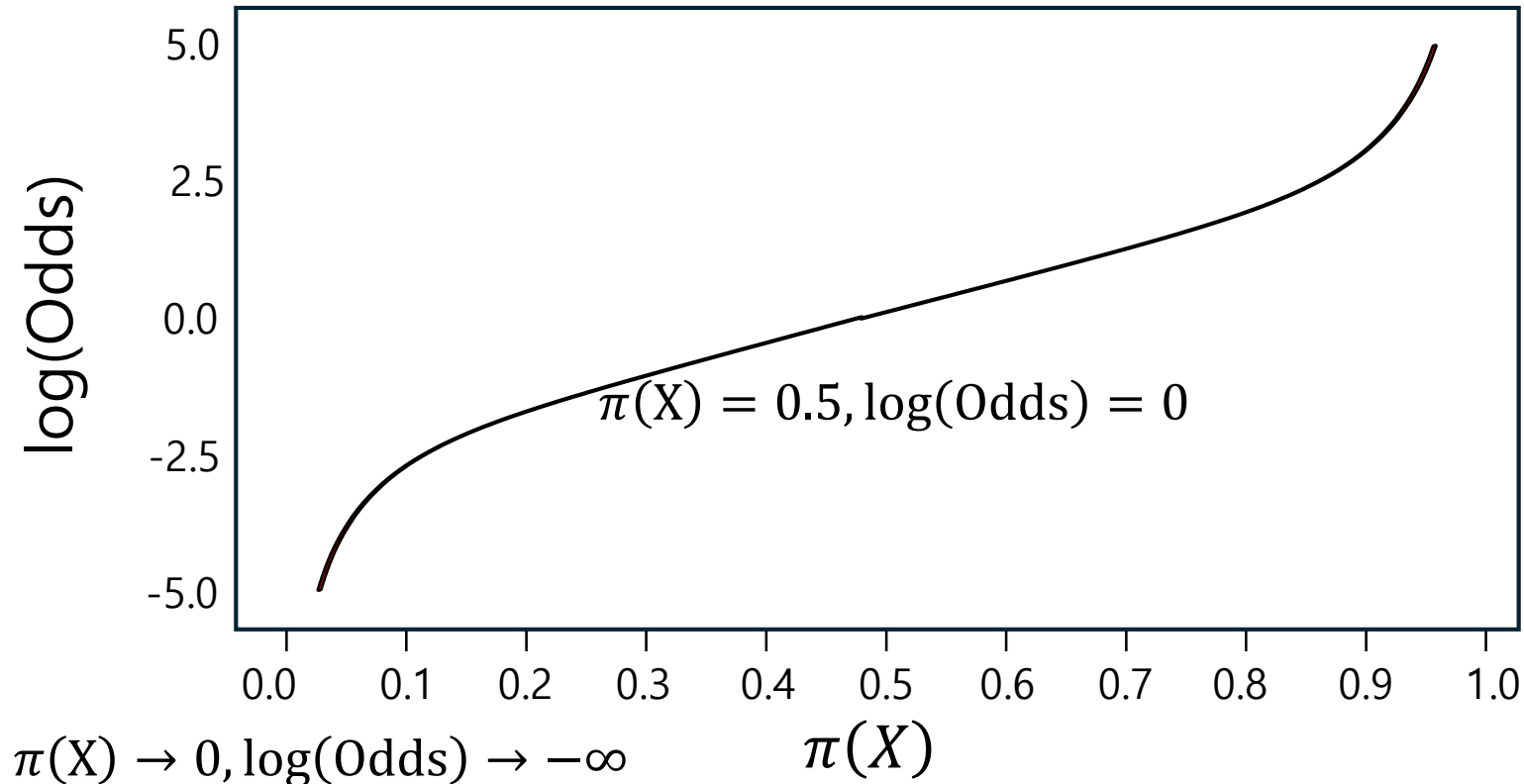
$\beta_1$ 의 의미 : x가 한단위 증가했을 때 log(odds)의 증가량

# 로지스틱 회귀모델 - Odds

- 성공확률  $\pi(X)$ 에 따른  $\log(\text{Odds})$ 의 그래프

$$\log \frac{\pi(x)}{1 - \pi(x)}$$

$\pi(X) \rightarrow 1, \log(\text{Odds}) \rightarrow \infty$



# 다중 로지스틱 회귀모델

- 다중 로지스틱 회귀모델 (입력변수 X가 2개 이상)

$$\pi(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)}}$$

# 다중 로지스틱 회귀모델

- 다중 로지스틱 회귀모델 (입력변수 X가 2개 이상)

$$\pi(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)}}$$

$$Odds = \frac{\pi(X=x)}{1 - \pi(X=x)} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}$$

# 다중 로지스틱 회귀모델

- 다중 로지스틱 회귀모델 (입력변수 X가 2개 이상)

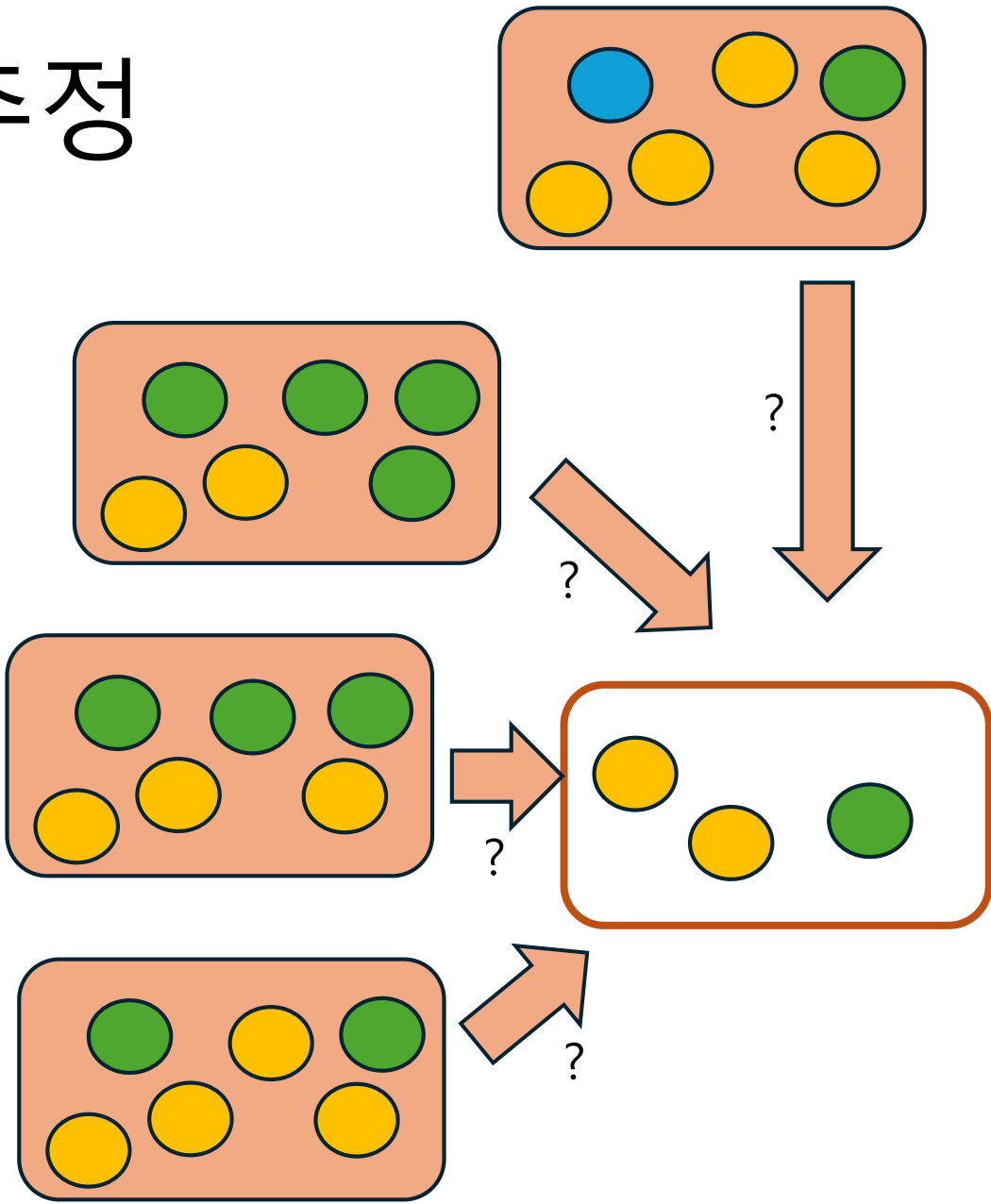
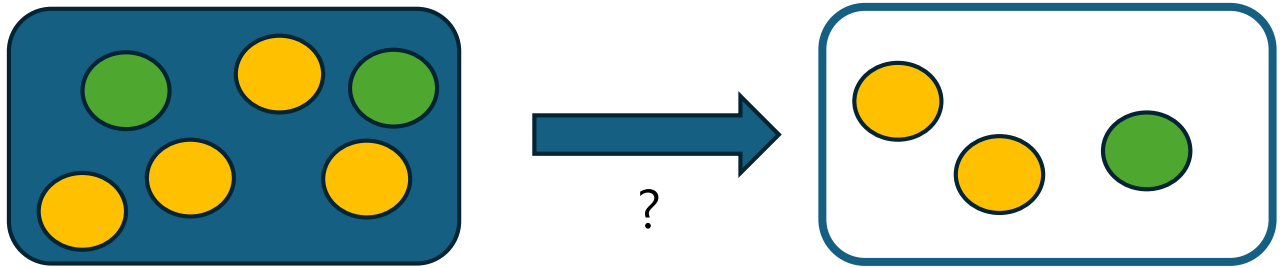
$$\pi(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)}}$$

$$Odds = \frac{\pi(X=x)}{1 - \pi(X=x)} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}$$

$$\log(Odds) = \log\left(\frac{\pi(X)}{1 - \pi(X)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

# 파라미터 추정

- 최대 우도 추정법 (Maximum Likelihood Estimation)
- 확률: 어떤 고정된 분포에서 특정 값이 나올 확률
- 우도: 어떤 고정된 값이 특정 분포에서 나왔을 확률





# 파라미터 추정

- 로지스틱 회귀모델 학습: 최대 우도 추정법 (Maximum Likelihood Estimation)

$$f_i(y_i) = \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i}, i = 1, 2, \dots, n$$

$$P(y_i = 1) = \pi_i$$

$$P(y_i = 0) = 1 - \pi_i$$

$$L = \prod_i f_i(y_i) = \prod_i \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i}$$

# 파라미터 추정

- 로지스틱 회귀모델 학습: 최대 우도 추정법 (Maximum Likelihood Estimation)

$$f_i(y_i) = \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i}, i = 1, 2, \dots, n$$

$$P(y_i = 1) = \pi_i$$

$$P(y_i = 0) = 1 - \pi_i$$

$$L = \prod_i f_i(y_i) = \prod_i \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i}$$

$$\begin{aligned} \ln L &= \ln \left[ \prod_i \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i} \right] \\ &= \ln \prod_i \left[ \frac{\pi(x_i)}{1 - \pi(x_i)} \right]^{y_i} + \sum_i \ln(1 - \pi(x_i)) \end{aligned}$$

# 파라미터 추정

- 로지스틱 회귀모델 학습: 최대 우도 추정법 (Maximum Likelihood Estimation)

$$f_i(y_i) = \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i}, i = 1, 2, \dots, n$$

$$P(y_i = 1) = \pi_i$$

$$P(y_i = 0) = 1 - \pi_i$$

$$L = \prod_i f_i(y_i) = \prod_i \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i}$$

$$\begin{aligned} \ln L &= \ln \left[ \prod_i \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i} \right] \\ &= \ln \prod_i \left[ \frac{\pi(x_i)}{1 - \pi(x_i)} \right]^{y_i} + \sum_i \ln(1 - \pi(x_i)) \\ &= \sum_i y_i \ln \left[ \frac{\pi(x_i)}{1 - \pi(x_i)} \right] + \sum_i \ln(1 - \pi(x_i)) \end{aligned}$$

# 파라미터 추정

- 로지스틱 회귀모델 학습: 최대 우도 추정법 (Maximum Likelihood Estimation)

$$f_i(y_i) = \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i}, i = 1, 2, \dots, n$$

$$P(y_i = 1) = \pi_i$$

$$P(y_i = 0) = 1 - \pi_i$$

$$L = \prod_i f_i(y_i) = \prod_i \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i}$$

$$\ln L = \ln \left[ \prod_i \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i} \right]$$

$$= \ln \prod_i \left[ \frac{\pi(x_i)}{1 - \pi(x_i)} \right]^{y_i} + \sum_i \ln(1 - \pi(x_i))$$

$$= \sum_i y_i \ln \left[ \frac{\pi(x_i)}{1 - \pi(x_i)} \right] + \sum_i \ln(1 - \pi(x_i))$$

$$= \sum_i y_i (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p) + \sum_i \ln(1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p})$$

# 파라미터 추정

- 로지스틱 회귀모델 학습: 최대 우도 측정법 (Maximum Likelihood Estimation)

$$\ln L = \sum_i y_i (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p) + \sum_i \ln(1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p})$$

- 위 로그-우도함수 (log likelihood function)가 최대가 되는 파라미터  $\beta$  결정
- 로그-우도함수 (log likelihood function)는 파라미터  $\beta$  에 대해 비선형이므로 선형회귀 모델과 같이 명시적인 해가 존재하지 않음 (No closed-form solution exists)
- Iterative reweight least square, Conjugate gradient, Newton's method 등의 수치 최적화 알고리즘을 이용하여 해를 구함

# 파라미터 추정

- Cross entropy: 두 확률분포  $(p(x), q(x))$ 의 차이

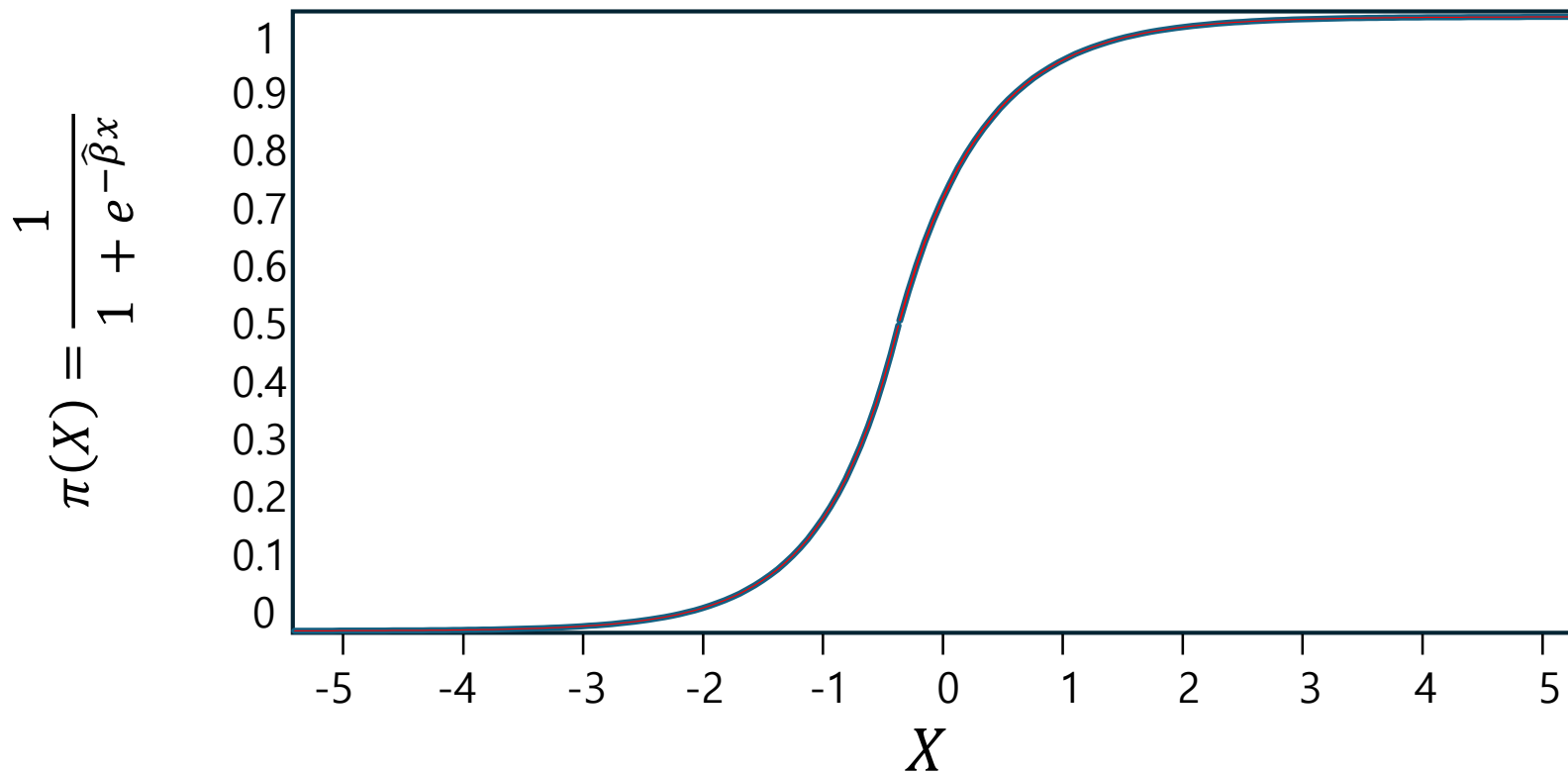
$$\text{Cross entropy} = -\sum p(x) \log q(x)$$

- Cross entropy: 음의 log likelihood function의 기대값
- Log likelihood function을 최대 = 입력 분포  $p(x)$ 와 파라미터가 주어졌을 때, 출력 분포  $q(x)$ 의 확률을 최대
- Cross entropy를 최소 = 입력 분포  $p(x)$ 와 출력 분포  $q(x)$ 의 차이를 최소
- Log likelihood function을 최대 = cross entropy를 최소

# 로지스틱 회귀모델 - 결과 및 해석

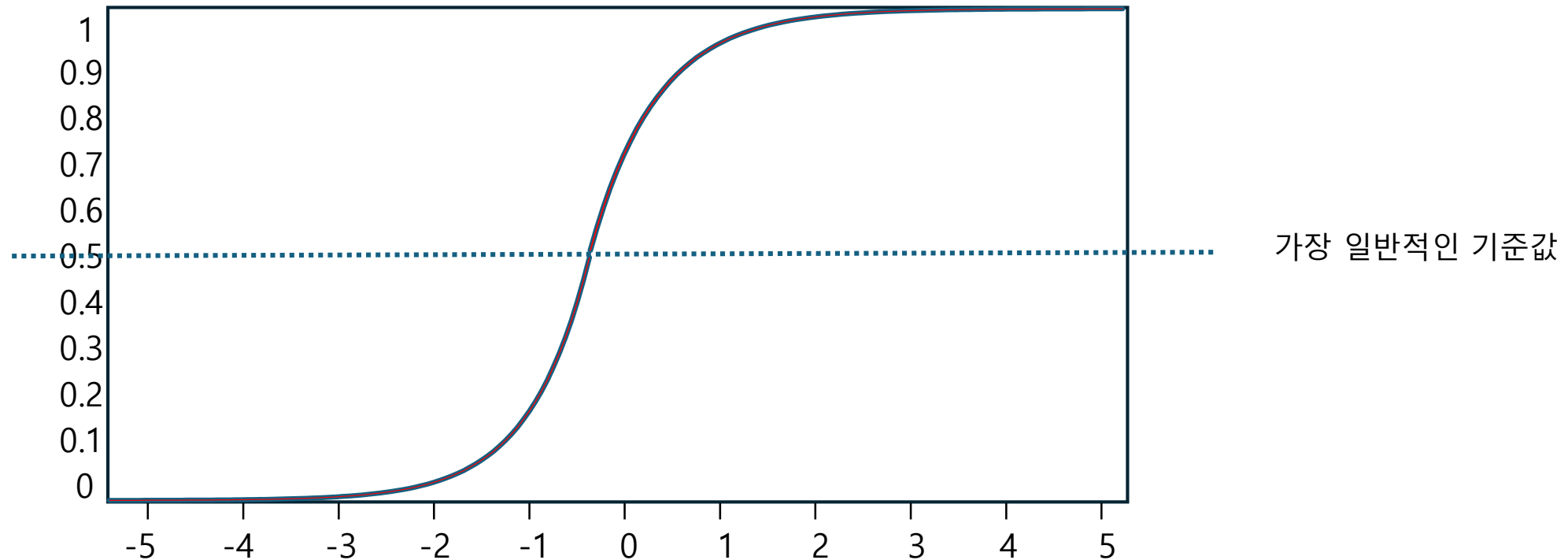
- 파라미터가 추정되고 난 이후 최종모델

$$\pi(X) = f(X) = \frac{1}{1 + e^{-(\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \dots + \widehat{\beta}_p x_p)}} = \frac{1}{1 + e^{-\widehat{\beta}x}}$$



# 로지스틱 회귀모델 - 결과 및 해석

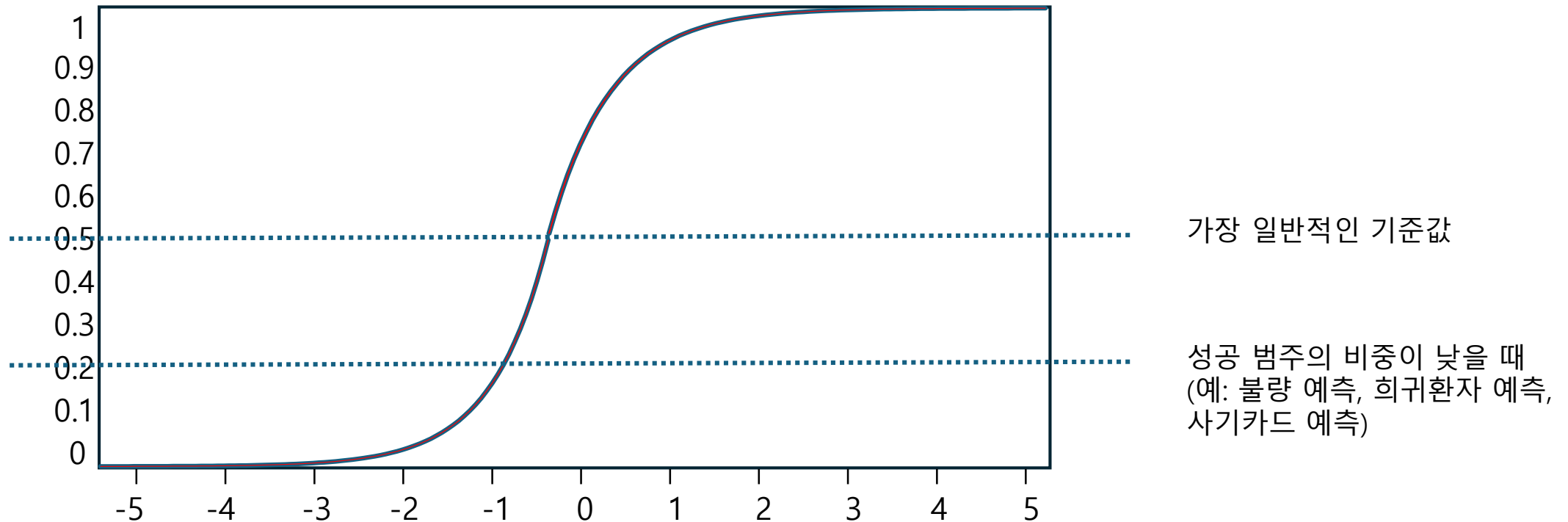
- 이진 분류를 위한 기준값(threshold) 설정
  - 일반적으로 0.5 사용





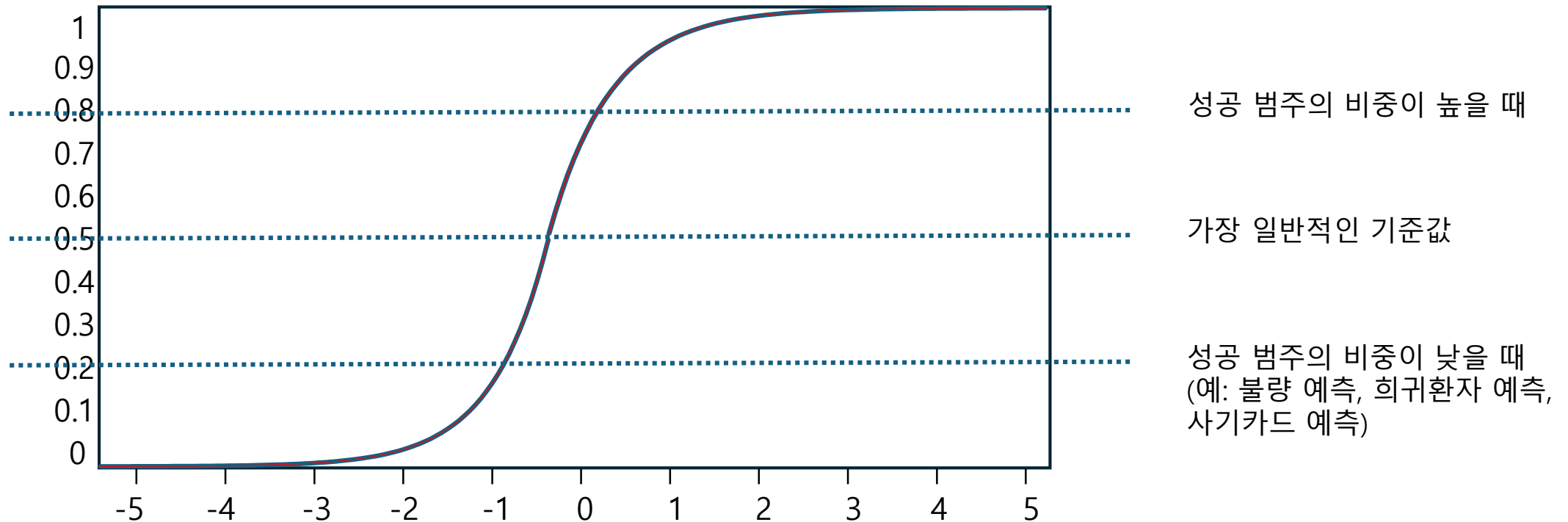
# 로지스틱 회귀모델 - 결과 및 해석

- 이진 분류를 위한 기준값(threshold) 설정
  - 일반적으로 0.5 사용



# 로지스틱 회귀모델 - 결과 및 해석

- 이진 분류를 위한 기준값(threshold) 설정
  - 일반적으로 0.5 사용



# 로지스틱 회귀모델 - 결과 및 해석

- (1) 선형회귀모델

$$f(X) = \widehat{\beta}_0 + \widehat{\beta}_1x_1 + \widehat{\beta}_2x_2 + \cdots + \widehat{\beta}_px_p$$

입력 변수가 1단위 증가할 때 출력변수의 변화량

- (2) 로지스틱 회귀모델

$$\log(Odds) = \widehat{\beta}_0 + \widehat{\beta}_1x_1 + \widehat{\beta}_2x_2 + \cdots + \widehat{\beta}_px_p$$

입력변수가 1단위 증가할 때 로그오드의 변화량

# 로지스틱 회귀모델 - 결과 및 해석

- 승산 비율: Odds Ratio

$$\frac{\text{odds}(x_1+1, x_2, \dots, x_n)}{\text{odds}(x_1, x_2, \dots, x_n)} = \frac{e^{\widehat{\beta}_0 + \widehat{\beta}_1(x_1+1) + \widehat{\beta}_2x_2 + \dots + \widehat{\beta}_px_p}}{e^{\widehat{\beta}_0 + \widehat{\beta}_1x_1 + \widehat{\beta}_2x_2 + \dots + \widehat{\beta}_px_p}} = e^{\beta_1}$$

- 나머지 입력변수는 모두 고정시킨 상태에서 한 변수를 1단위 증가시켰을 때 변화하는 Odds의 비율
- X1이 1단위 증가하면 성공에 대한 승산 비율이  $e^{\beta_1}$  만큼 변화함
- 회귀 계수가 양수 -> 성공확률 증가 (성공확률  $\geq 1$ )
- 회귀 계수가 음수 -> 성공확률 감소 ( $0 \leq$  성공확률  $< 1$ )

# 로지스틱 회귀모델 - 예제

- 로지스틱 회귀분석 결과 및 해석 (대출 여부를 확인하는 데이터)

$$f(X) = \frac{1}{1 + e^{-(\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \dots + \widehat{\beta}_{12} x_{12})}}$$

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-13.20165825	2.46772742	0.00000009	*
Age	-0.04453737	0.09096102	0.62439483	0.95643985
Experience	0.05657264	0.09005365	0.5298661	1.05820346
Income	0.0657607	0.00422134	0	1.06797111
Family	0.57155931	0.10119002	0.00000002	1.77102649
CCAvg	0.18724874	0.06153848	0.00234395	1.20592725
Mortgage	0.00175308	0.00080375	0.02917421	1.00175464

# 로지스틱 회귀모델 - 결과 및 해석

- Coefficient (로지스틱 회귀계수, 추정된 파라미터 값)
  - 해당 변수가 1단위 증가할 때 로그오드의 변화량
  - 양수이면 성공확률과 양의 상관관계, 음수이면 성공 확률과 음의 상관관계

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-13.20165825	2.46772742	0.00000009	*
Age	-0.04453737	0.09096102	0.62439483	0.95643985
Experience	0.05657264	0.09005365	0.5298661	1.05820346
Income	0.0657607	0.00422134	0	1.06797111
Family	0.57155931	0.10119002	0.00000002	1.77102649
CCAvg	0.18724874	0.06153848	0.00234395	1.20592725
Mortgage	0.00175308	0.00080375	0.02917421	1.00175464

# 로지스틱 회귀모델 - 결과 및 해석

- Std. Error (추정 파라미터의 표준편차)
  - 추정 파라미터의 신뢰구간 (구간추정)을 구축할 때 사용

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-13.20165825	2.46772742	0.00000009	*
Age	-0.04453737	0.09096102	0.62439483	0.95643985
Experience	0.05657264	0.09005365	0.5298661	1.05820346
Income	0.0657607	0.00422134	0	1.06797111
Family	0.57155931	0.10119002	0.00000002	1.77102649
CCAvg	0.18724874	0.06153848	0.00234395	1.20592725
Mortgage	0.00175308	0.00080375	0.02917421	1.00175464

# 로지스틱 회귀모델 - 결과 및 해석

- p-value
  - 해당 변수가 통계적으로 유의미한지 여부를 알려주는 지표
  - 해당 파라미터 값이 0인지 여부를 통계적으로 판단 (가설검정)

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-13.20165825	2.46772742	0.00000009	*
Age	-0.04453737	0.09096102	0.62439483	0.95643985
Experience	0.05657264	0.09005365	0.5298661	1.05820346
Income	0.0657607	0.00422134	0	1.06797111
Family	0.57155931	0.10119002	0.00000002	1.77102649
CCAvg	0.18724874	0.06153848	0.00234395	1.20592725
Mortgage	0.00175308	0.00080375	0.02917421	1.00175464



# 로지스틱 회귀모델 - 결과 및 해석

- Odds (Odds Ratio)
  - 나머지 입력변수는 모두 고정시킨 상태에서 한 변수를 1단위 증가시켰을 때 변화하는 Odds(성공확률)의 비율
  - Experience = 1.058 -> 경험이 1년 더 많으면 대출 확률이 1.058배 증가

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-13.20165825	2.46772742	0.00000009	*
Age	-0.04453737	0.09096102	0.62439483	0.95643985
Experience	0.05657264	0.09005365	0.5298661	1.05820346
Income	0.0657607	0.00422134	0	1.06797111
Family	0.57155931	0.10119002	0.00000002	1.77102649
CCAvg	0.18724874	0.06153848	0.00234395	1.20592725
Mortgage	0.00175308	0.00080375	0.02917421	1.00175464

# 로지스틱 회귀모델 예제

- 나이, 사회적 지위, 거주지역과 질병유무와의 관계
- 사회적 지위는 원래 3개의 범주 (상, 중, 하)를 갖는 변수  
-> 2개의 이진변수 ( $X_2, X_3$ )로 표현 (상-> (0,0), 중->(1,0), 하->(0,1))
- 거주지역은 2개 범주 (지역1 -> 0, 지역2 -> 1)

	Age	Socioeconomic Status		City Sector	Disease Status
	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$
1	33	0	0	0	0
2	35	0	0	0	0
3	6	0	0	0	0
4	60	0	0	1	0
5	18	0	1	1	1
6	26	0	1	0	0
...	...	...	...	...	...
98	35	0	0	0	1

# 로지스틱 회귀모델 예제

$$f(x) = \frac{1}{1 + e^{-(-2.31 + 0.03X_1 + 0.41X_2 - 0.31X_3 + 1.57X_4)}}$$

	Age	Socioeconomic Status		City Sector	Disease Status
	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$
1	33	0	0	0	0
2	35	0	0	0	0
3	6	0	0	0	0
4	60	0	0	1	0
5	18	0	1	1	1
6	26	0	1	0	0
...	...	...	...	...	...
98	35	0	0	0	1

# 로지스틱 회귀모델 예제

$$f(x) = \frac{1}{1 + e^{-(-2.31 + 0.03X_1 + 0.41X_2 - 0.31X_3 + 1.57X_4)}}$$

	Age	Socioeconomic Status		City Sector	Disease Status
	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$
1	33	0	0	0	0
2	35	0	0	0	0
3	6	0	0	0	0
4	60	0	0	1	0
5	18	0	1	1	1
6	26	0	1	0	0
...	...	...	...	...	...
98	35	0	0	0	1
99	46	1	0	1	??

# 로지스틱 회귀모델 예제

$$f(x) = \frac{1}{1 + e^{-(-2.31 + 0.03X_1 + 0.41X_2 - 0.31X_3 + 1.57X_4)}}$$

	Age	Socioeconomic Status		City Sector	Disease Status
	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$
1	33	0	0	0	0
2	35	0	0	0	0
3	6	0	0	0	0
4	60	0	0	1	0
5	18	0	1	1	1
6	26	0	1	0	0
...	...	...	...	...	...
98	35	0	0	0	1
<b>99</b>	<b>46</b>	<b>1</b>	<b>0</b>	<b>1</b>	<b>0.74</b>

# 로지스틱 회귀모델 예제

$$f(x) = \frac{1}{1 + e^{-(-2.31 + 0.03X_1 + 0.41X_2 - 0.31X_3 + 1.57X_4)}}$$

	Age	Socioeconomic Status		City Sector	Disease Status
	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$
1	33	0	0	0	0
2	35	0	0	0	0
3	6	0	0	0	0
4	60	0	0	1	0
5	18	0	1	1	1
6	26	0	1	0	0
...	...	...	...	...	...
98	35	0	0	0	1
99	46	1	0	1	1

# 로지스틱 회귀모델 예제

Regression Coefficient	Odds Ratio
$\beta_1$	1.030
$\beta_2$	1.505
$\beta_3$	0.737
$\beta_4$	4.829

- $\beta_1$ 의 odds ratio = 1.030 -> 나이가 1살 증가하면 질병 걸릴 확률 1.03배 증가
- $\beta_4$ 의 odds ratio = 4.829 -> 거주지역이 2이면 질병 걸릴 확률 4.829배 증가

Thank you