로지스틱회귀 (Logistic Regression) 모델

로지스틱 회귀모델 강의자료 개요

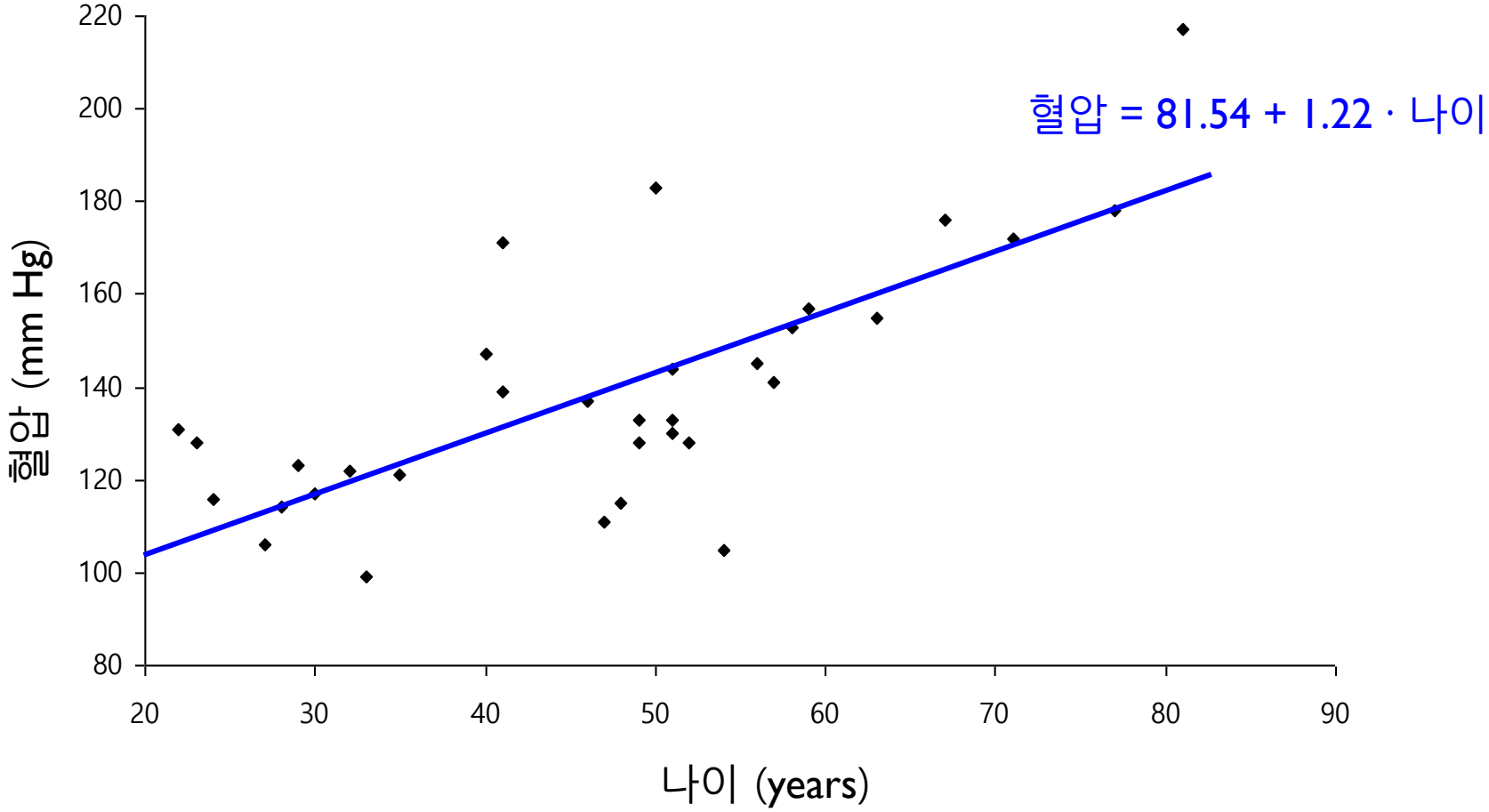
- 로지스틱 회귀모델 배경
- 로지스틱 회귀모델 형태
- 아드 (Odds)
- 파라미터 추정
- 로지스틱 회귀모델 결과 및 해석
- 로지스틱 회귀모델 예제

로지스틱 회귀모델 배경

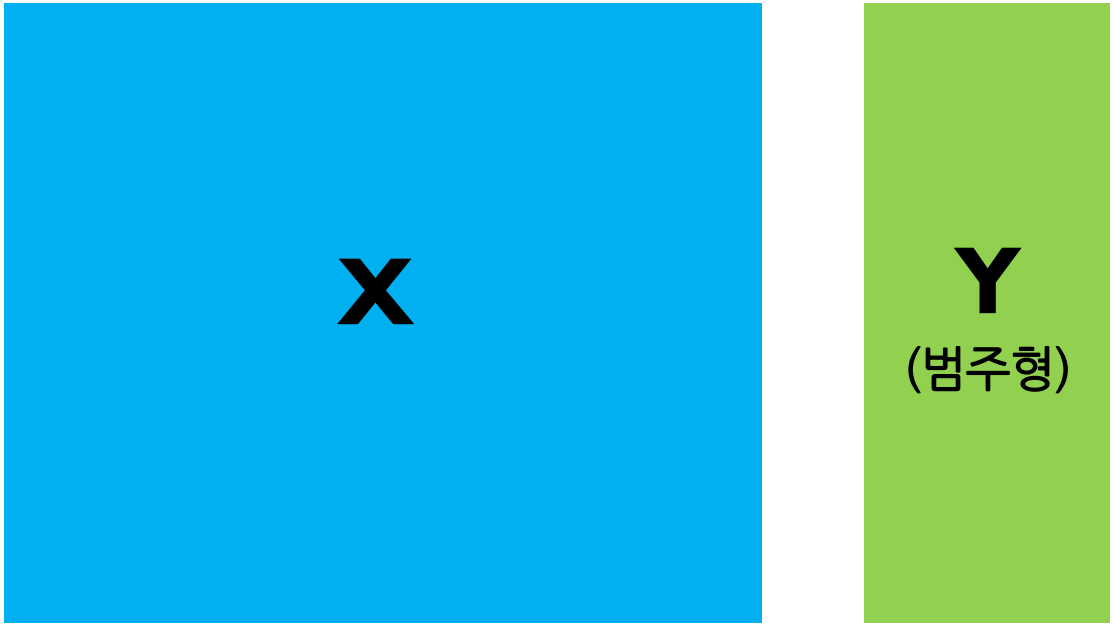
- 33명의 나이와 혈압 사이의 관계

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


혈압 = 81.54 + 1.22 · 나이
220
200
180
160
140
120
100
80
혈압 (mm Hg)
20
30
40
50
60
70
80
90
나이 (years)

로지스틱 회귀모델 필요성



- **범주형 반응변수**
  - 이진변수 (반응변수 값 ∈ 0 or 1)
  - 멀티변수 (반응변수 값 ∈ 1 or 2 or 3 이상)
- 선형회귀모델과는 다른 방식으로 접근해야 될 필요성

로지스틱 회귀모델 필요성

- **로지스틱 회귀모델 사용**
  - 새로운 관측치가 왔을 때 이를 기존 범주 중 하나로 예측 (범주예측)
- **응용예제**
  - 제품이 불량인지 양품인지 분류
  - 고객이 이탈고객인지 잔류고객인지 분류
  - 카드 거래가 정상인지 사기인지 분류
  - 내원 고객이 질병이 있는지 없는지 분류
  - 특정인의 유전자 정보를 보고 백혈병의 유무
  - 이메일이 스팸인지 정상메일인지
  - 페이스북 피드에서 보이게 할지 숨길지

로지스틱 회귀모델 이론 배경

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

Assume $E(\varepsilon_i) = 0$ $E(Y_i) = \beta_0 + \beta_1 X_i$

Consider Y_i to be a Bernoulli 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

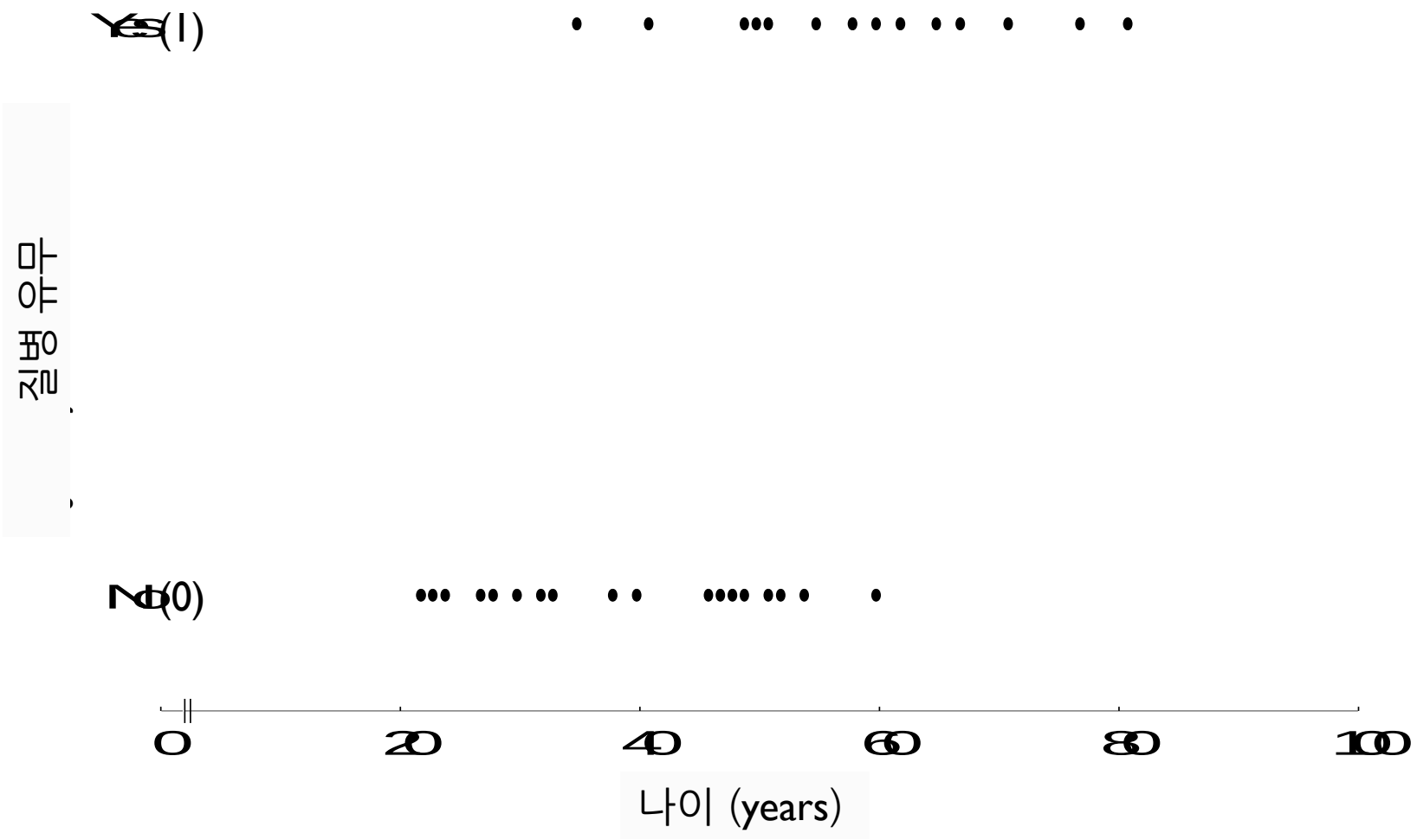
로지스틱 회귀모델

- 출력변수가 연속형이 아닌 이진범주형 질병유무?

나이	질병유무	나이	질병유무	나이	질병유무
22	0	40	0	54	0
23	0	41	1	55	1
24	0	46	0	58	1
27	0	47	0	60	1
28	0	48	0	60	0
30	0	49	1	62	1
30	0	49	0	65	1
32	0	50	1	67	1
33	0	51	0	71	1
35	1	51	1	77	1
38	0	52	0	81	1

로지스틱 회귀모델

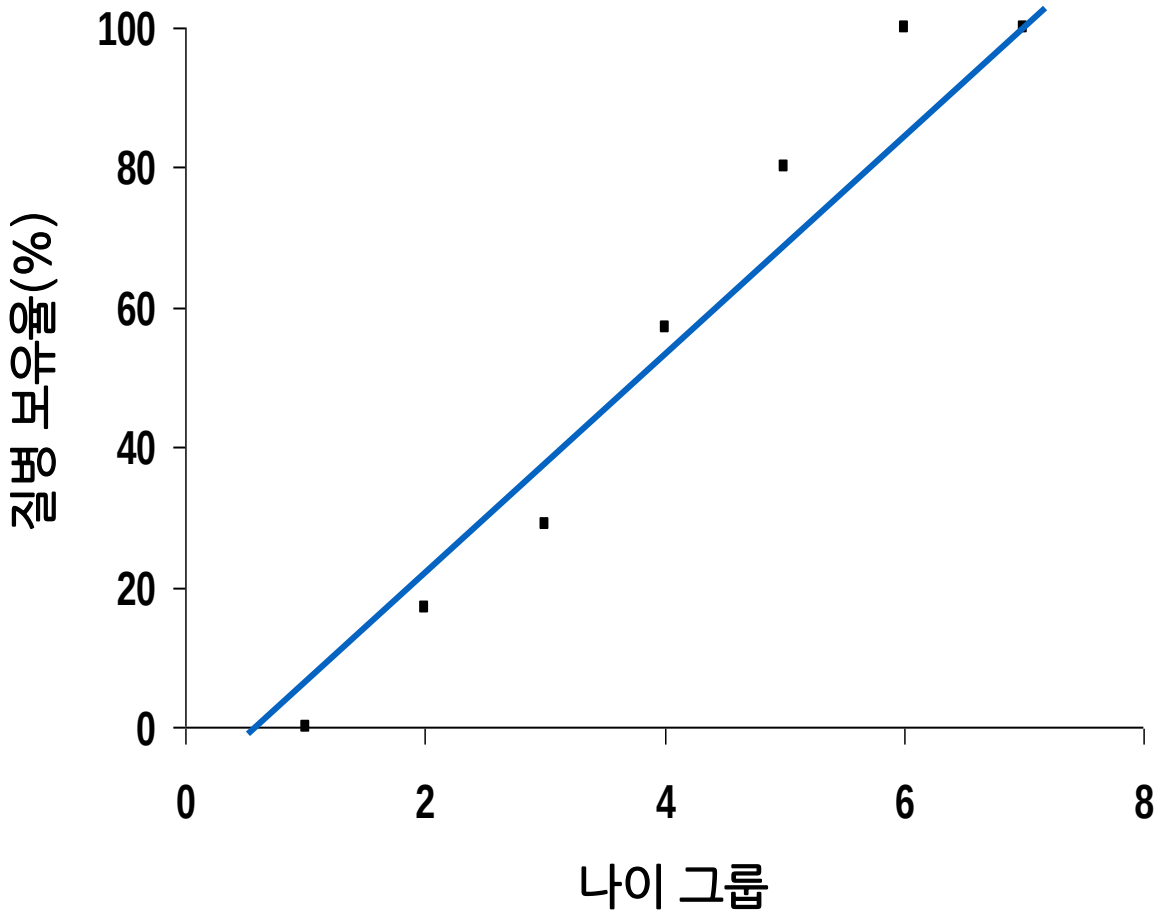
- 두 변수 사이의 관계식은 선형??



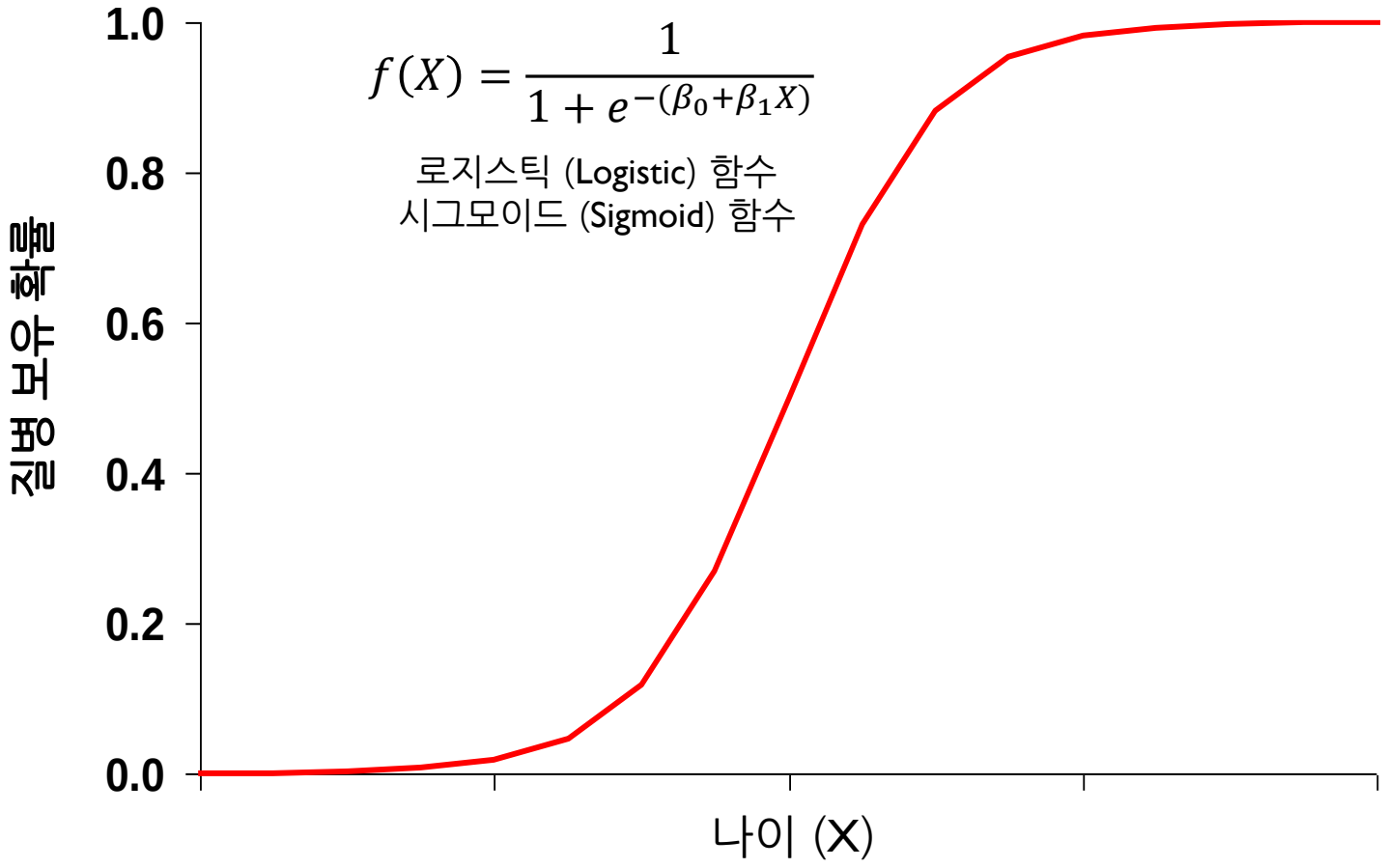
로지스틱 회귀모델

		질병	
나이 그룹	그룹내 수	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

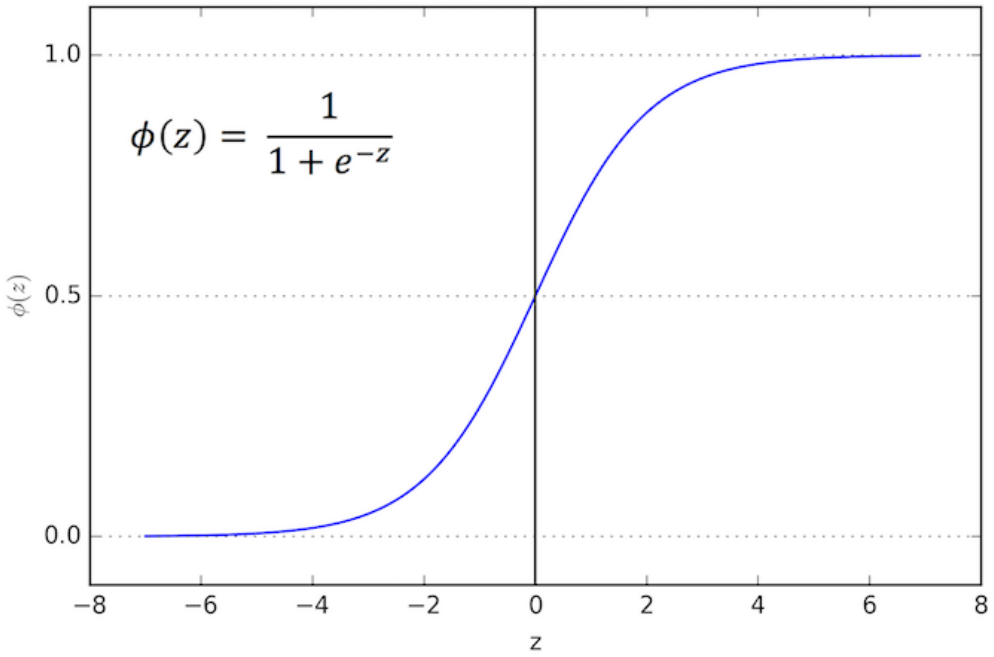
로지스틱 회귀분석 알고리즘



로지스틱 회귀분석 알고리즘 – 로지스틱 함수



로지스틱 회귀분석 알고리즘 – 로지스틱 함수



- Logistic function, Sigmoid function, Squashing function (Large input → Small output)
- 아웃풋 범위: 0~1
- 인풋값에 대해 단조증가 (혹은 단조감소) 함수
- 미분결과를 아웃풋의 함수로 표현 가능 (Gradient learning method에 유용하게 사용)

$$\frac{d\emptyset(z)}{dz} = \frac{1}{1+e^{-z}}\left(1 - \frac{1}{1+e^{-z}}\right) = \emptyset(z)(1- \emptyset(z))$$

로지스틱 회귀분석 알고리즘 – 로지스틱 함수

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

단순로지스틱 회귀모델: 입력변수 X가 1개인 로지스틱 회귀모델

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

관측치 x가 범주 1에 속할 확률

(Probability that an observation x belongs to class 1)

로지스틱 회귀모델 - β_1의 해석

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

β_1의 해석 → 직관적이지 못함

- 승산 (Odds)
  - 성공 확률을 p 로 정의할 때, 실패 대비 성공 확률 비율

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

$$p = 1 \rightarrow odd = \infty$$
$$p = 0 \rightarrow odd = 0$$

로지스틱 회귀모델 - β_1의 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
96101 GERMANY	5/1	4/1	96117 SWEDEN	80/1	100/1
96102 BRAZIL	8/1	5/1	96118 SERBIA	100/1	150/1
96103 FRANCE	10/1	11/2	96119 SENEGAL	500/1	150/1
96104 SPAIN	8/1	13/2	96120 EGYPT	500/1	200/1
96105 ARGENTINA	8/1	7/1	96121 ICELAND	1000/1	200/1
96106 BELGIUM	15/1	10/1	96122 PERU	500/1	200/1
96107 ENGLAND	25/1	15/1	96123 NIGERIA	150/1	200/1
96108 PORTUGAL	30/1	20/1	96124 JAPAN	150/1	300/1
96109 URUGUAY	50/1	30/1	96125 COSTA RICA	200/1	250/1
96110 COLOMBIA	20/1	40/1	96126 AUSTRALIA	300/1	300/1
96111 RUSSIA	20/1	40/1	96127 MOROCCO	500/1	500/1
96112 CROATIA	60/1	40/1	96128 IRAN	2000/1	500/1
96113 POLAND	100/1	40/1	96129 SOUTH KOREA	200/1	500/1
96114 MEXICO	50/1	60/1	96130 TUNISIA	500/1	1000/1
96115 DENMARK	100/1	100/1	96131 PANAMA	1000/1	1000/1
96116 SWITZERLAND	80/1	100/1	96132 SAUDI ARABIA	1000/1	1000/1

Odds current as of 1/22/18

로지스틱 회귀모델 - β_1의 해석 - Odds

11 : 2

프랑스의 우승 odds는 2/11

프랑스의 우승 확률은 2/13 = 0.15 (15 %)

500 : 1

대한민국의 우승 odds는 1/500

대한민국의 우승확률은 1/501 ≒ 0.001996 (0.1996%)

로지스틱 회귀모델 - β_1의 해석 - Odds

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

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

Odds: 범주 0에 속할 확률 대비 범주 1에 속할 확률

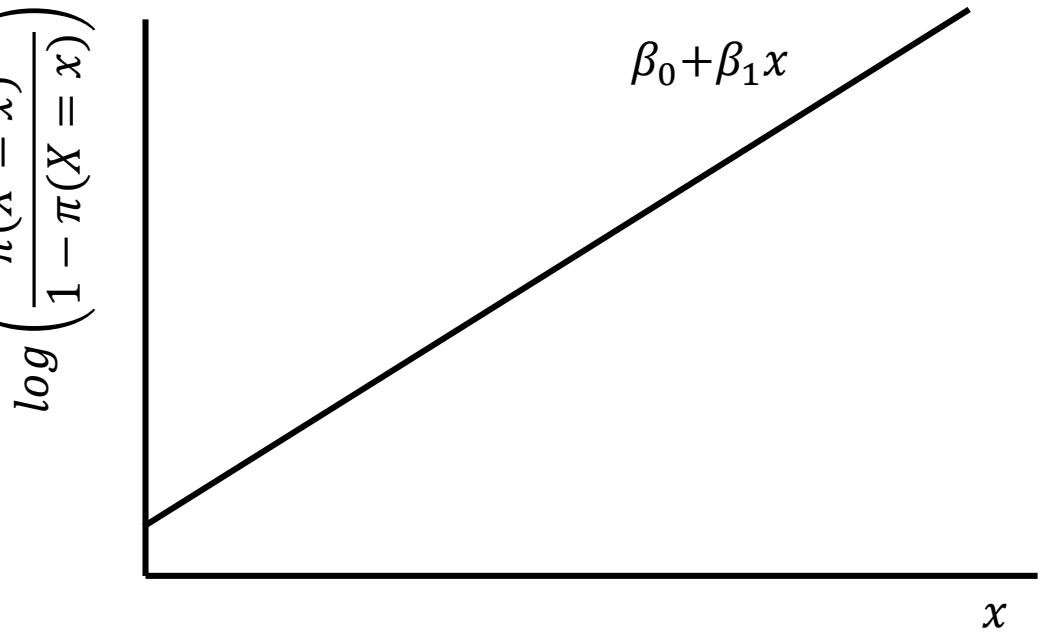
(The ratio of the probability of belonging to class 1 to the probability of belonging to class 0)

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

Logit Transform (로짓 변환)

로지스틱 회귀모델 - β_1의 해석 - Odds

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

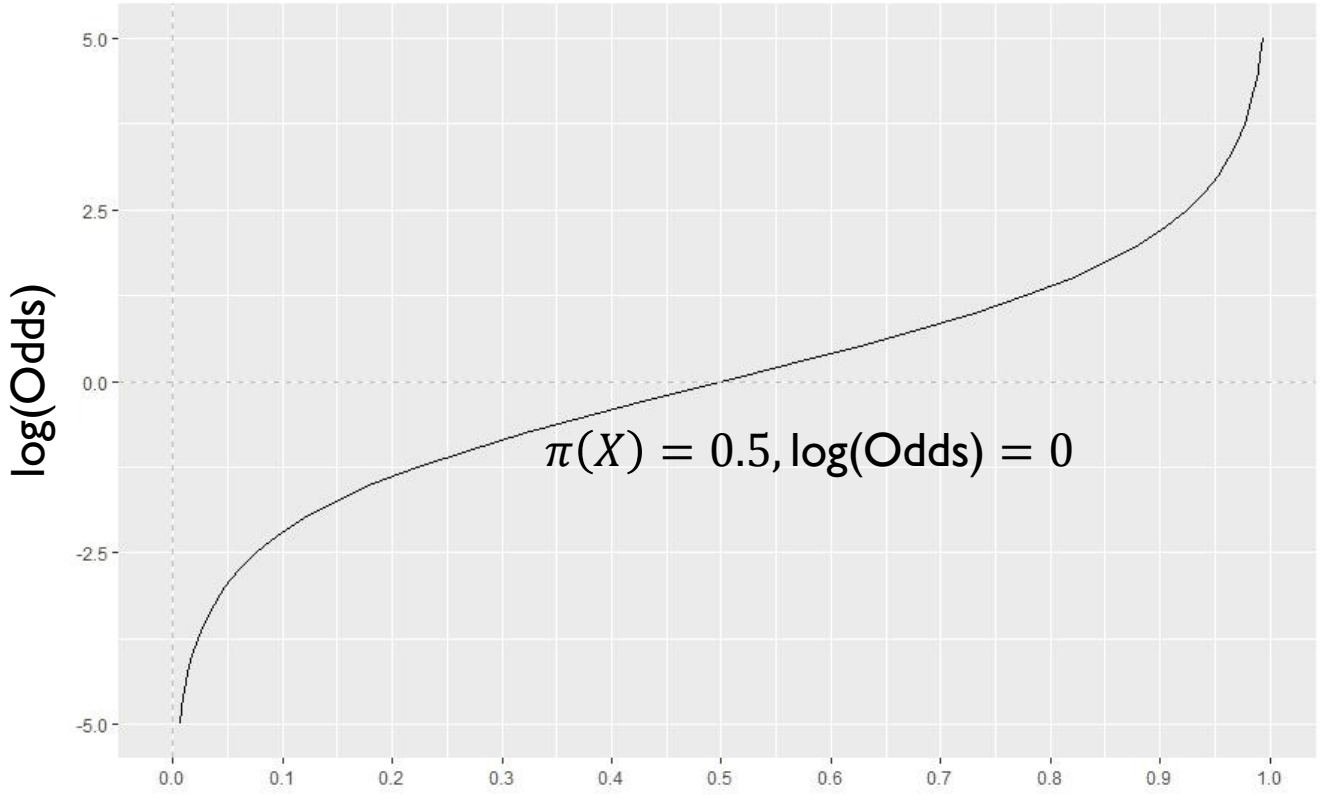


β_1의 의미: x가 한단위 증가 했을 때 log(odds)의 증가량

로지스틱 회귀모델 - Odds

- 성공 확률 $\pi(X)$ 에 따른 log(Odds)의 그래프

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



$\pi(X) \to 0, \log(\text{Odds}) \to -\infty$

파라미터 추정

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

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

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

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

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

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

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

$$= \sum_i y_i\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\right) - \sum_i \ln(1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p})$$

파라미터 추정

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

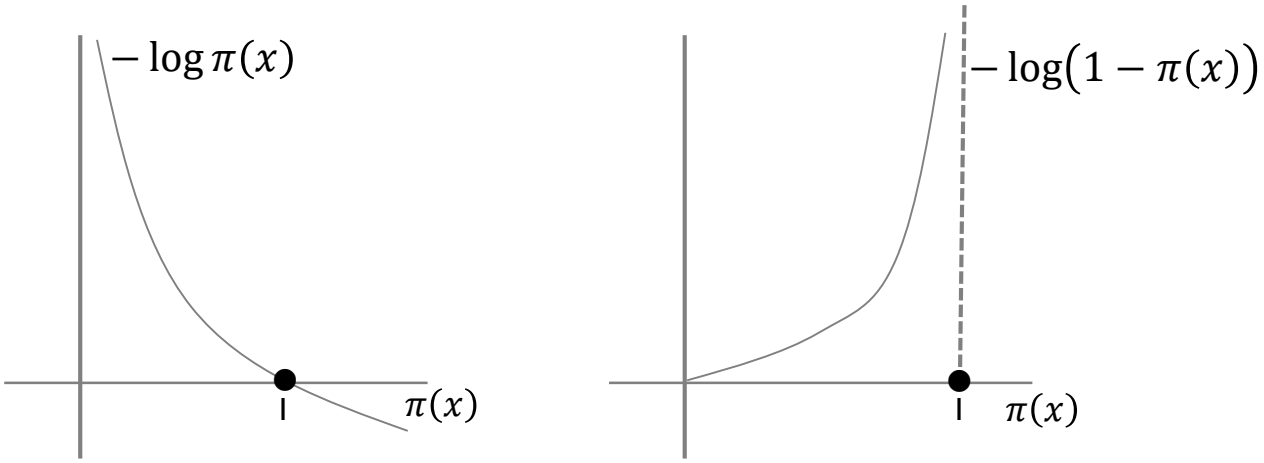
$$\ln L = \sum_i y_i\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\right) + \sum_i \ln\left(1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}\right)$$

  - 위 로그-우도함수 (**log likelihood function**)가 최대가 되는 파라미터 β 결정
  - 로그-우도함수 (**log likelihood function**)는 파라미터 β에 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

$$C(\pi(x), y) = \begin{cases} -\log \pi(x), & y = 1 \\ -\log(1 - \pi(x)), & y = 0 \end{cases}$$



$$C(\pi(x), y) = -y \log \pi(x) - (1 - y) \log(1 - \pi(x))$$

$$\min_{\beta} C(\pi(x), y)$$

파라미터 추정

- Cross entropy: 두 확률분포 $(p(x), q(x))$의 차이 $(p(x) = 0\ or\ 1, q(x) = \pi(x)\ or\ 1 - \pi(x))$

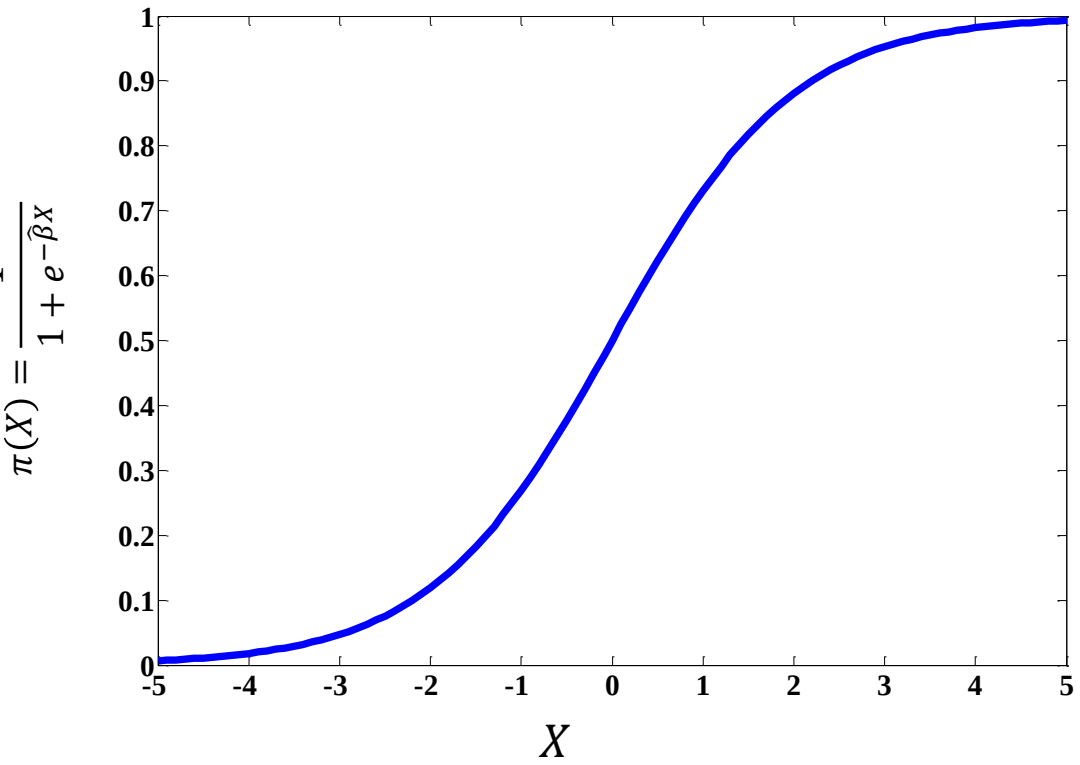
$$\text{Cross entropy} = -\sum p(x) \log q(x)$$

- 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 출력분포 $q(x)$의 차이를 최소
- Log likelihood function을 최대 = cross entropy를 최소

로지스틱 회귀모델 - 결과 및 해석

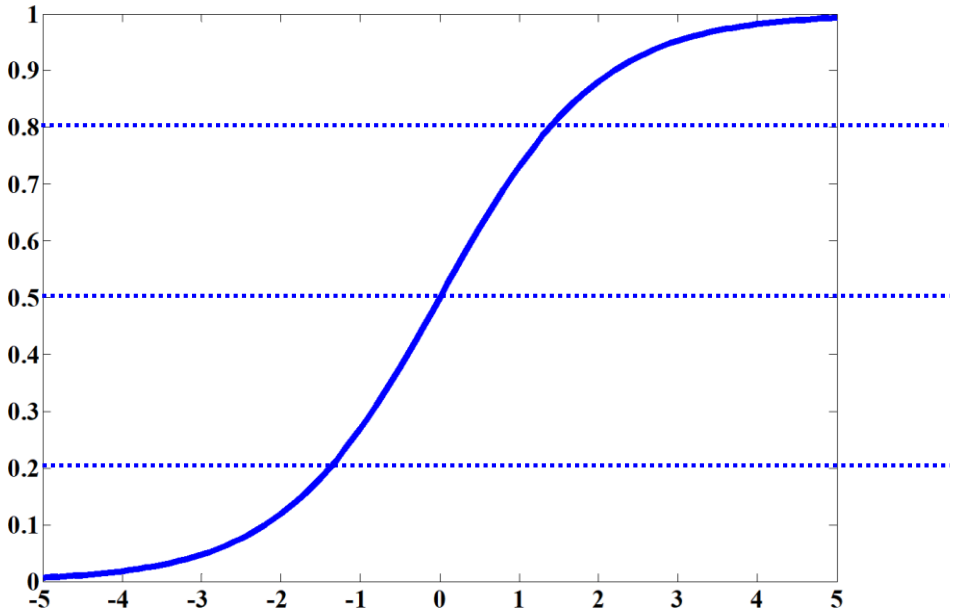
❖ 파라미터가 추정되고 난 이후 최종모델

$$\pi(X) = f(X) = \frac{1}{1 + e^{-\left(\widehat{\beta_0} + \widehat{\beta_1} X_1 + \cdots + \widehat{\beta_p} X_p\right)}} = \frac{1}{1 + e^{-\widehat{\beta} X}}$$



로지스틱 회귀모델 - 결과 및 해석

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



성공 범주의 비중이 높을 때

가장 일반적인 기준값

성공 범주의 비중이 낮을 때
(예: 불량 예측, 희귀환자 예측, 사기카드예측)

로지스틱 회귀모델 - 결과 및 해석

(1) 선형회귀모델

$$f(X) = \widehat{\beta_0} + \widehat{\beta_1} X_1 + \cdots + \widehat{\beta_p} X_p$$

입력변수가 1단위 증가할 때 **출력변수의 변화량**

(2) 로지스틱회귀모델

$$log(Odds) = \widehat{\beta_0} + \widehat{\beta_1} X_1 + \cdots + \widehat{\beta_p} X_p$$

입력변수가 1단위 증가할 때 **로그아드의 변화량**

로지스틱 회귀모델 - 결과 및 해석

- 승산 비율: Odds Ratio

$$\frac{odds(x_1+1, x_2, \cdots, x_n)}{odds(x_1, x_2, \cdots, x_n)} = \frac{e^{\widehat{\beta_0}+\widehat{\beta_1}(X_1+1)+\cdots+\widehat{\beta_p}X_p}}{e^{\widehat{\beta_0}+\widehat{\beta_1}X_1+\cdots+\widehat{\beta_p}X_p}} = e^{\widehat{\beta_1}}$$

- 나머지 입력변수는 모두 고정시킨 상태에서 한 변수를 1단위 증가시켰을 때 변화하는 Odds의 비율
- x_1이 1단위 증가하면 성공에 대한 승산 비율이 e^{β_1} 만큼 변화함
- 회귀 계수가 양수 → 성공확률 증가 (성공확률 ≥ 1)
- 회귀 계수가 음수 → 성공확률 감소 ($0 \leq$ 성공확률 < 1)

로지스틱 회귀모델 - 예제

- 로지스틱 회귀분석 결과 및 해석 (대출 여부를 예측하는 데이터)

$$f(X) = \frac{1}{1 + e^{-(\widehat{\beta_0} + \widehat{\beta_1} X_1 + \cdots + \widehat{\beta_{12}} X_{12})}}$$

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
Securities Account	-0.85484785	0.41863668	0.04115349	0.42534789
CD Account	3.46900773	0.44893095	0	32.10486984
Online	-0.84355801	0.22832377	0.00022026	0.43017724
CreditCard	-0.96406376	0.28254223	0.00064463	0.38134006
EducGrad	4.58909273	0.38708162	0	98.40509796
EducProf	4.52272701	0.38425466	0	92.08635712

로지스틱 회귀모델 - 결과 및 해석

- **Coefficient (로지스틱 회귀계수, 추정된 파라미터 값)**
  - 해당 변수가 1단위 증가할 때 로그아드의 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
Securities Account	-0.85484785	0.41863668	0.04115349	0.42534789
CD Account	3.46900773	0.44893095	0	32.10486984
Online	-0.84355801	0.22832377	0.00022026	0.43017724
CreditCard	-0.96406376	0.28254223	0.00064463	0.38134006
EducGrad	4.58909273	0.38708162	0	98.40509796
EducProf	4.52272701	0.38425466	0	92.08635712

로지스틱 회귀모델 - 결과 및 해석

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

Input variables	Coefficien	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
Securities Account	-0.85484785	0.41863668	0.04115349	0.42534789
CD Account	3.46900773	0.44893095	0	32.10486984
Online	-0.84355801	0.22832377	0.00022026	0.43017724
CreditCard	-0.96406376	0.28254223	0.00064463	0.38134006
EducGrad	4.58909273	0.38708162	0	98.40509796
EducProf	4.52272701	0.38425466	0	92.08635712

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
Securities Account	-0.85484785	0.41863668	0.04115349	0.42534789
CD Account	3.46900773	0.44893095	0	32.10486984
Online	-0.84355801	0.22832377	0.00022026	0.43017724
CreditCard	-0.96406376	0.28254223	0.00064463	0.38134006
EducGrad	4.58909273	0.38708162	0	98.40509796
EducProf	4.52272701	0.38425466	0	92.08635712

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	0.0000000	*
Age	-0.04453737	0.09096102	0.6243948	0.95643985
Experience	0.05657264	0.09005365	0.529866	1.05820346
Income	0.0657607	0.00422134	0	1.06797111
Family	0.57155931	0.10119002	0.0000000	1.77102649
CCAvg	0.18724874	0.06153848	0.0023439	1.20592725
Mortgage	0.00175308	0.00080375	0.029174	1.00175464
Securities Account	-0.85484785	0.41863668	0.0411534	0.42534789
CD Account	3.46900773	0.44893095	0	32.10486984
Online	-0.84355801	0.22832377	0.0002202	0.43017724
CreditCard	-0.96406376	0.28254223	0.0006446	0.38134006
EducGrad	4.58909273	0.38708162	0	98.40509796
EducProf	4.52272701	0.38425466	0	92.08635712

로지스틱 회귀모델 예제

- 나이, 사회적 지위, 거주지역과 질병유무와의 관계
- 사회적 지위는 원래 3개의 범주 (상, 중, 하)를 갖는 변수

  → 2개의 이진변수 (X_1, X_2)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
99	**46**	**1**	**0**	**1**	**1**

로지스틱 회귀모델 예제

Regression Coefficient	Odds Ratio
β_1	1.030
β_2	1.505
β_3	0.737
β_4	4.829

- β_1의 odds ratio = 1.030 → 나이가 1살 증가하면 질병 걸릴 확률 1.03배 증가
- β_4의 odds ratio = 4.829 → 거주지역이 2이면 질병 걸릴 확률 4.829배 증가

로지스틱 회귀모델 강의자료 개요

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

EOD