전통적 머신러닝 알고리즘 심화

[이론] 김지수 [실습] 강민현, 김준목

{gisookim, mkang, jm0917}@unist.ac.kr

2023년 9월 13일

강의 목차

벡터와 행렬 연산

벡터와 행렬

- **배열** (array) : 동일한 특성을 가지는 요소(예: 숫자)가 특정 규칙으로 나열되어 있는 데이터 집합
- **벡터** (vector) : 1차원의 배열.

$$a = [a_1, a_2, \cdots, a_n]^T \in \mathbb{R}^n$$

a_i: 벡터 a 의 i 번째 원소.

- **행렬** (matrix) : 2차원의 (사각형 모양의) 배열. 크기 $n \times m$ 인 행렬은 n 개의 행(row)과 m 개의 열(column)을 가짐.

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1m} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nm} \end{pmatrix} \in \mathbb{R}^{n \times m}$$

a_{ij}: 행렬 A의 i 번째 행, j 번째 열에 위치한 원소.
만약 $n = m$이면 A를 정사각행렬이라 부름.
각 행은 행벡터, 각 열은 열벡터라 부르기도 함.

벡터와 행렬 연산

1. 덧셈 & 뺄셈

$$\begin{pmatrix} 2 & 5 & -1 \\ 1 & 0 & -3 \end{pmatrix} + \begin{pmatrix} -7 & 3 & 1 \\ 4 & 1 & 0 \end{pmatrix} = \begin{pmatrix} -5 & 8 & 0 \\ 5 & 1 & -3 \end{pmatrix}.$$

: 덧셈 & 뺄셈은 크기가 같은 행렬에 대해서만 가능. 원소별로 수행.

2. 상수 곱

$$3 \times \begin{pmatrix} 2 & 5 & -1 \\ 1 & 0 & -3 \end{pmatrix} = \begin{pmatrix} 6 & 15 & -3 \\ 3 & 0 & -9 \end{pmatrix}$$

3. 벡터의 내적

$$a \circ b = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n = \sum_{i=1}^{n} a_i b_i$$

벡터와 행렬 연산

4. 행렬의 곱
$n \times m$ 행렬 A 와 $m \times p$ 행렬 B를 곱하면 $n \times p$ 행렬이 됨:

$$A \times B = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix} \times \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mp} \end{pmatrix}$$

$$= \begin{pmatrix} a_{1\cdot} \circ b_{\cdot 1} & a_{1\cdot} \circ b_{\cdot 2} & \cdots & a_{1\cdot} \circ b_{\cdot p} \\ a_{2\cdot} \circ b_{\cdot 1} & a_{2\cdot} \circ b_{\cdot 2} & \cdots & a_{2\cdot} \circ b_{\cdot p} \\ a_{3\cdot} \circ b_{\cdot 1} & a_{3\cdot} \circ b_{\cdot 2} & \cdots & a_{3\cdot} \circ b_{\cdot p} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n\cdot} \circ b_{\cdot 1} & a_{n\cdot} \circ b_{\cdot 2} & \cdots & a_{n\cdot} \circ b_{\cdot p} \end{pmatrix} (\neq B \times A)$$

example

$$\begin{pmatrix} 2 & 5 & -1 \\ 1 & 0 & -3 \end{pmatrix} \times \begin{pmatrix} 3 \\ 1 \\ 5 \end{pmatrix} = \begin{pmatrix} 6 \\ -12 \end{pmatrix}$$

행렬 곱셈은 (왼쪽 행렬의 열의 개수)=(오른쪽 행렬의 행의 개수) 일때만 가능.

벡터와 행렬 연산

5. 행렬 곱의 성질

$$A(BC) = (AB)C, \quad (A+B)C = AC + BC.$$

6. 행렬의 전치(Transposition)

$$\begin{pmatrix} 2 & 5 & -1 \\ 1 & 0 & -3 \end{pmatrix}^T = \begin{pmatrix} 2 & 1 \\ 5 & 0 \\ -1 & -3 \end{pmatrix}$$

- 전치는 행렬의 행을 열로 바꿔줌 (=열을 행으로 바꿔줌)
- $(A^T)_{ij} = A_{ji}$
- $n \times m$ 행렬의 전치는 $m \times n$ 행렬임.
- $(AB)^T = B^T A^T$
- 두 벡터 a 와 b 의 내적 $a \circ b$ 를 $a^T b$ 로 표현할 수 있음.

특수한 행렬들

1. **단위 행렬 (Identity matrix)**: 대각 원소가 모두 1이고 비대각 원소가 모두 0인 정사각 행렬.

$$I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad I_n = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

<u>**성질**</u> 임의의 $n \times m$ 행렬 A에 대하여, $A \times I_m = I_n \times A = A$ 이 성립함.

2. **대칭 행렬 (Symmetric matrix)**: 대각원소들을 기준으로 대칭을 이루는 행렬. 즉, 정사각 행렬이면서 $A = A^T$를 만족하는 행렬 A.
<u>example</u>

$$A = \begin{pmatrix} 3 & 2 & 1 \\ 2 & -5 & 7 \\ 1 & 7 & -4 \end{pmatrix} = A^T$$

특수한 행렬들

3. $(n \times n)$ 정사각 행렬 A 에 대하여 아래 식을 만족하는 행렬 B가 존재하면 A를 가역(invertible) 행렬이라 함.

$$A \times B = B \times A = I_n$$

B는 A의 역행렬이며 $(B = A^{-1})$, A도 B의 역행렬임 $(A = B^{-1})$.

Notes

- 가역행렬이 아닌 행렬도 존재한다. 예를 들어, $A = \begin{pmatrix} 1 & 0 \\ 3 & 0 \end{pmatrix}$ 는 가역행렬이 아님.
- 2×2 행렬의 역행렬을 구하는 식:

$$\text{If } A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad A^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

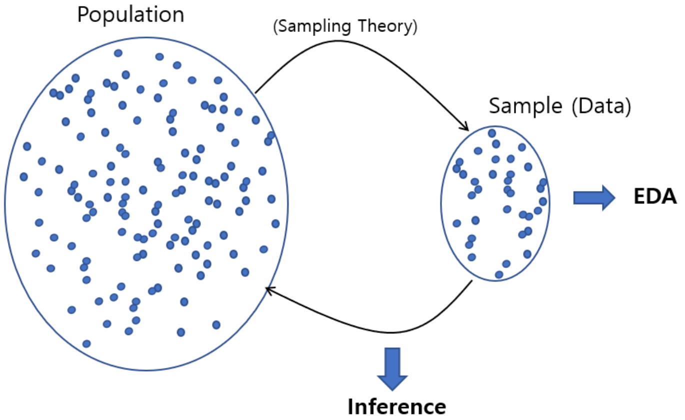
2×2 행렬 A는 $ad - bc \neq 0$인 경우에만 가역행렬임.

간단한 통계량

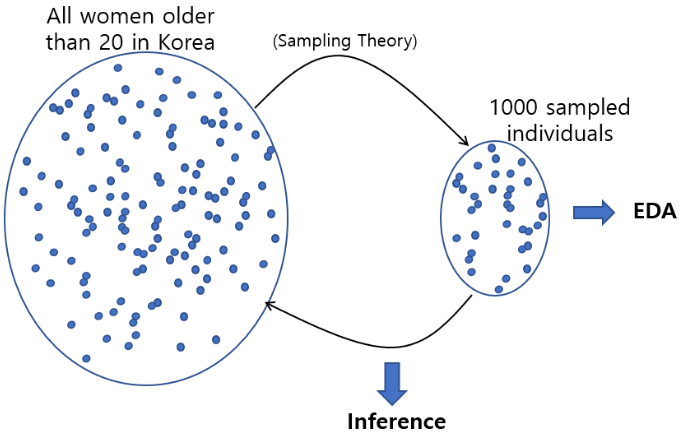
통계?

- **통계학**은 관심 모집단으로부터 데이터를 수집하고, 데이터를 과학적으로 분석하고, 데이터로 부터 모집단에 대한 확률론적 추론을 하는 방법론을 통칭함.
- 통계학의 두 갈래:
  - ▷ **탐색적 데이터 분석** (Exploratory Data Analysis; EDA)
  - ▷ **통계적 추론** (Statistical Inference)

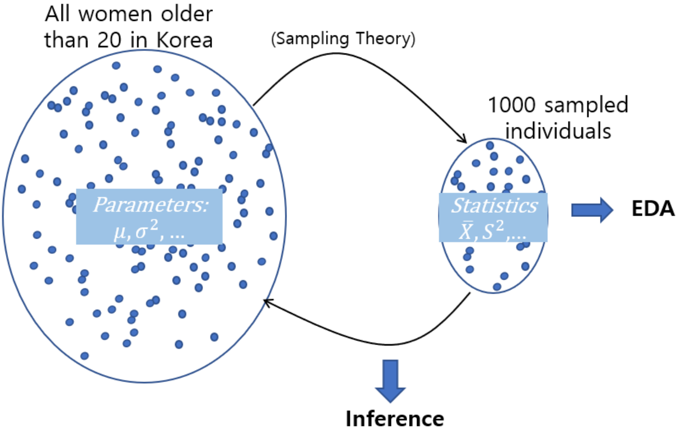
통계?



통계?


All women older than 20 in Korea
(Sampling Theory)
1000 sampled individuals
EDA
Inference

통계?



example: 한국의 모든 20세 이상 여성의 평균 키(μ)에 대한 추론을 하기 위하여 1000명을 추출하고 그들의 평균 키($\bar{X}$)를 계산함.

중심 경향

- 평균(mean)

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- 중앙값(median)
  -변수들을 크기 순서대로 나열한 후 계산: $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$.

$$\check{X} = \frac{1}{2}\{X_{(\lfloor \frac{n+1}{2} \rfloor)} + X_{(\lceil \frac{n+1}{2} \rceil)}\}$$

어떤 통계량이 더 좋은가?

중심 경향

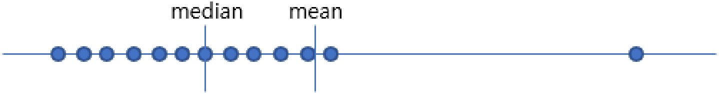
- 평균(mean)

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- 중앙값(median)
  -변수들을 크기 순서대로 나열한 후 계산: $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$.

$$\check{X} = \frac{1}{2}\{X_{(\lfloor \frac{n+1}{2} \rfloor)} + X_{(\lceil \frac{n+1}{2} \rceil)}\}$$

어떤 통계량이 더 좋은가?



산포

- 표본 분산

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

  S를 표준편차라 한다.
- 엔트로피

$$\epsilon(X) = -\sum_{c\in\mathcal{C}} \hat{p}_c \log_2 \hat{p}_c.$$

  ($\hat{p}_c$: 데이터 내에서 범주 c의 비율.)
  - $0 \leq \epsilon(X) \leq \log_2(|\mathcal{C}|)$.
  - $\epsilon(X) = 0 \Leftrightarrow$ 특정 범주 c에 대하여 $\hat{p}_c = 1$.
  - $\epsilon(X) = \log_2(|\mathcal{C}|) \Leftrightarrow$ 모든 범주 c에 대하여 $\hat{p}_c = 1/|\mathcal{C}|$.
- 그 외 통계량들 : 분위수, 사분위수(Q1: 25% 분위수, Q2: 중앙값, Q3: 75% 분위수), 사분위수 범위(Inter Quartile Range; IQR; Q3-Q1)

공변 (Covariation)

- 표본 공분산

$$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$$

- 표본 상관계수

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}$$

  - 두 변수의 선형 관계를 수치화
  - $|r| = 1$는 완벽한 선형관계를 의미함
  - $r = 0$는 선형 관계가 없음을 의미
  - 부호는 선형관계의 방향을 의미 (양/음의 상관관계)
  - $r_{(ax+b)(cy+d)} = r_{xy}$ $(a, b, c, d \in \mathbb{R}$ and $ac > 0)$

Supplementary: 공분산과 상관계수

공분산 (S_{xy})	**상관계수 (r_{xy})**
두 확률 변수가 공변하는 양상을 측정하는 척도	두 확률변수의 선형 관계를 측정하는 척도
단위에 따라 다름	단위가 없음. 단위의 변화에 영향을 받지 않음.
$-\infty \leq S_{xy} \leq \infty$	$-1 \leq r_{xy} \leq 1$

Table: 공분산과 상관계수

Supplementary: 공분산과 상관계수

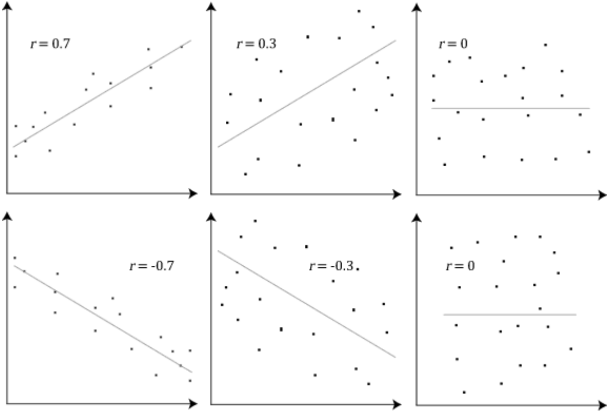


Figure: Retrieved from https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

통계량의 행렬 연산

평균과 분산, 공분산을 행렬 연산으로 나타내면?

- 평균

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\mathbf{1} \circ X = \frac{1}{n}\mathbf{1}^T X$$

$\mathbf{1} = [1, 1, \cdots, 1]^T$, $X = [X_1, X_2, \cdots, X_n]^T$

- 분산

퀴즈!

- 공분산

퀴즈!

확률 분포와 통계적 추론

확률 분포

모집단(population)은 특정 확률 분포를 따름. 확률 분포는 확률 변수의 값에 따른 확률의 분포를 가리킴.

- 이산형 변수의 **확률 질량 함수** $p(x)$: $p(x) = P(X = x)$
  (1) $0 \le p(x) \le 1$
  (2) $\sum_{x \in \mathcal{X}} p(x) = 1$.
  (3) $P(a \le X \le b) = \sum_{a \le x \le b} p(x)$.
- 연속형 변수의 **확률 밀도 함수** $p(x)$: $\int_a^b p(x)dx = P(a \le X \le b)$
  (1) $0 \le p(x)$
  (2) $\int_{-\infty}^{\infty} p(x)dx = 1$.
  (3) $P(a \le X \le b) = P(a < X \le b) = P(a \le X < b) = P(a < X < b) = \int_a^b p(x)dx$.

모집단의 평균

$$E(X) = \begin{cases} \sum_{x \in \mathcal{X}} xp(x) \\ \int_{-\infty}^{\infty} xp(x)dx \end{cases}$$

$$E(f(X)) = \begin{cases} \sum_{x \in \mathcal{X}} f(x)p(x) \\ \int_{-\infty}^{\infty} f(x)p(x)dx \end{cases}$$

Properties

- 임의의 $a, b \in \mathbb{R}$에 대하여 $E(aX + b) = aE(X) + b$.
- $E(af_1(X) + bf_2(X)) = aE(f_1(X)) + bE(f_2(X))$.

모집단의 분산

Let $\mu = E(X)$.

$$Var(X) = E[(X-\mu)^2] = \begin{cases} \sum_{x \in \mathcal{X}} (x-\mu)^2 p(x) \\ \int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx \end{cases} = E(X^2) - \mu^2$$

$sd(X) = \sqrt{Var(X)}$

Properties

- 임의의 $a, b \in \mathbb{R}$에 대하여, $Var(aX+b) = a^2 Var(X)$.

결합 확률 분포

- **두** 이산형 확률변수의 확률 질량 함수: $p(x,y) = P(X = x, Y = y)$
  (1) $0 \leq p(x,y) \leq 1$
  (2) $\sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} p(x,y) = 1.$
  (3) $P(a \leq X \leq b, c \leq Y \leq d) = \sum_{c \leq y \leq d} \sum_{a \leq x \leq b} p(x,y).$

- **두** 연속형 확률변수의 확률 밀도 함수:

$$\int_c^d \int_a^b p(x,y)dxdy = P(a \leq X \leq b, c \leq Y \leq d)$$

<u>주변 분포</u>

- $p_X(x) = \sum_{y \in \mathcal{Y}} p(x,y).$
- $p_X(x) = \int_{-\infty}^{\infty} p(x,y)dy$

결합 확률 분포

기대값

$$E(f(X,Y)) = \begin{cases} \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} f(x,y)p(x,y) \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)p(x,y)dxdy \end{cases}$$

두 확률변수의 **독립**

임의의 값 $x \in \mathcal{X}$ 와 $y \in \mathcal{Y}$에 대하여,

$$p(x,y) = p(x)p(y),$$

가 성립하면 X와 Y는 독립이다.
두 확률변수 X와 Y가 독립이면,

$$E(XY) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} xyp(x,y) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} xyp(x)p(y)$$
$$= \sum_{x\in\mathcal{X}} xp(x) \sum_{y\in\mathcal{Y}} yp(y) = E(X)E(Y).$$

모집단의 공분산과 상관계수

$\mu_X = E(X)$, $\mu_Y = E(Y)$, $\sigma_X = sd(X)$, $\sigma_Y = sd(Y)$라 하자.

$$Cov(X, Y) = E\Big[(X - \mu_X)(Y - \mu_Y)\Big].$$

$$Corr(X, Y) = E\left[\frac{(X - \mu_X)}{\sigma_X}\frac{(Y - \mu_Y)}{\sigma_Y}\right].$$

Random sample의 개념

- **Random sample**은 동일한 분포에서 독립적으로 추출한 (identically and independently distributed; i.i.d.) 확률 변수 $X_1, X_2, \cdots, X_n$의 모음이다. 즉,
  1. X_1, X_2, ..., X_n 이 서로 독립이다.
  2. 임의의 $x \in \mathbb{R}$에 대하여 $p(X_1 = x) = p(X_2 = x) = \cdots = p(X_n = x)$.

특정 확률 분포를 가지는 모집단이 있다고 하자. 이 모집단으로 부터 무작위 반복추출로 표본을 추출하면 이 표본이 모집단의 random sample을 이룬다.

Random sample 을 이용하여 모집단에 대한 통계적 추론을 할 수 있다.

1. 베르누이 분포

베르누이 분포: $Ber(p)$ $(0 < p < 1)$

- 이항 변수의 분포.

x	0	1
$p(x) = P(X = x)$	$1 - p$	p

  예: 성공/실패, 동전 던질 때의 앞면/뒷면, 질병 진단에 대한 양성/음성 결과, 결혼 여부
- 확률 질량 함수: $p(x) = p^x(1-p)^{(1-x)}$.
- Properties:

$$
\begin{aligned}
E(X) &= 0 \times (1-p) + 1 \times p = p \\
Var(X) &= E\left[\{X - E(X)\}^2\right] = E(X^2) - \{E(X)\}^2 = p(1-p).
\end{aligned}
$$

1. 베르누이 분포

베르누이 분포 $Ber(p)$를 따르는 모집단으로부터 **random sample** $X_1, \cdots, X_n$을 추출하였다고 가정하자. 이 random sample 의 표본 평균은 다음을 만족함을 쉽게 확인 할 수 있다.

- $E(\bar{X}) = p$
- $Var(\bar{X}) = \frac{p(1-p)}{n}$

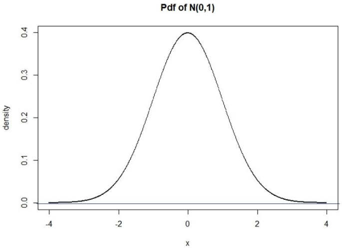
$\bar{X}$의 분포는 ??

$$
\begin{aligned}
P(n\bar{X} = m) &= P(X_1 + X_2 + \cdots + X_n = m) \\
&= \binom{n}{m} p^m (1-p)^{n-m} \\
&= \frac{n!}{m!(n-m)!} p^m (1-p)^{n-m}
\end{aligned}
$$

2. 정규분포

정규분포: $\mathcal{N}(\mu, \sigma^2)$.

- 연속형 변수의 분포.



- $x = \mu$에 대하여 대칭인 분포.
- 종 모양.
- 산포가 σ에 의해 결정됨.

- 확률 밀도 함수: $p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \ -\infty < x < \infty.$
- Properties:

$$E(X) = \int_{-\infty}^{\infty} xp(x)dx = \mu$$

$$Var(X) = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx = \sigma^2$$

2. 정규분포

- Properties:
  - 임의의 $a, b \in \mathbb{R}$에 대하여 $X \sim \mathcal{N}(\mu, \sigma^2)$이면,

$$aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2).$$

  - $X \sim \mathcal{N}(\mu, \sigma^2)$이면,

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1) : \text{표준 정규 분포}.$$

  - $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ 이며 X_1과 X_2 과 독립이면, 임의의 $a_1, a_2 \in \mathbb{R}$ 에 대하여,

$$a_1X_1 + a_2X_2 \sim \mathcal{N}(a_1\mu_1 + a_2\mu_2, a_1^2\sigma_1^2 + a_2^2\sigma_2^2).$$

$$\begin{aligned}
Var(a_1X_1 + a_2X_2) &= E\left[\{(a_1X_1 + a_2X_2) - E(a_1X_1 + a_2X_2)\}^2\right] \\
&= E\left[\{a_1(X_1 - \mu_1) + a_2(X_2 - \mu_2)\}^2\right] \\
&= E\left[a_1^2(X_1 - \mu_1)^2 + a_2^2(X_2 - \mu_2)^2 + 2a_1a_2(X_1 - \mu_1)(X_2 - \mu_2)\right] \\
&= a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + 0.
\end{aligned}$$

2. 정규분포

정규분포 $\mathcal{N}(\mu, \sigma^2)$를 따르는 모집단으로부터 **random sample** $X_1, \cdots, X_n$을 추출하였다고 가정하자. 이 random sample 의 표본 평균은 다음을 만족함을 쉽게 확인 할 수 있다.

- $E(\bar{X}) = \mu$
- $Var(\bar{X}) = \frac{\sigma^2}{n}$

$\bar{X}$의 분포는 ??

$$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}).$$

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

표본 평균의 성질

- 평균이 μ이고 분산이 σ^2인 임의의 모집단으로부터 **random sample** $X_1, \cdots, X_n$을 추출하였다고 가정하자. 표본 평균 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$의 성질은 아래와 같다.

$$E(\bar{X}) = E\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu.$$

$$\begin{aligned}
Var(\bar{X}) &= E\Big[\{\bar{X} - E(\bar{X})\}^2\Big] = E\Big[\{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\}^2\Big] = E\Big[\{\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n} n\mu\}^2\Big] \\
&= \frac{1}{n^2} E\Big[\{\sum_{i=1}^{n}(X_i - \mu)\}^2\Big] = \frac{1}{n^2} E\Big[\sum_{i=1}^{n}(X_i - \mu)^2 + \sum_{i \neq j}(X_i - \mu)(X_j - \mu)\Big] \\
&= \frac{1}{n^2}\Big[\sum_{i=1}^{n} E(X_i - \mu)^2 + \sum_{i \neq j} E\big[(X_i - \mu)(X_j - \mu)\big]\Big] \\
&= \frac{1}{n^2}\Big[n\sigma^2 + \sum_{i \neq j}\big[E(X_i - \mu)E(X_j - \mu)\big]\Big] = \frac{\sigma^2}{n}.
\end{aligned}$$

표본 평균의 성질

- 평균이 μ이고 분산이 σ^2인 임의의 모집단으로부터 **random sample** $X_1, \cdots, X_n$을 추출하였다고 가정하자. 표본 평균 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$의 성질은 아래와 같다.

$$E(\bar{X}) = E\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}n\mu = \mu.$$

$$\begin{aligned}
Var(\bar{X}) &= E\Big[\{\bar{X} - E(\bar{X})\}^2\Big] = E\Big[\{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\}^2\Big] = E\Big[\{\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}n\mu\}^2\Big] \\
&= \frac{1}{n^2}E\Big[\{\sum_{i=1}^{n}(X_i - \mu)\}^2\Big] = \frac{1}{n^2}E\Big[\sum_{i=1}^{n}(X_i - \mu)^2 + \sum_{i \neq j}(X_i - \mu)(X_j - \mu)\Big] \\
&= \frac{1}{n^2}\Big[\sum_{i=1}^{n} E(X_i - \mu)^2 + \sum_{i \neq j} E\big[(X_i - \mu)(X_j - \mu)\big]\Big] \\
&= \frac{1}{n^2}\Big[n\sigma^2 + \sum_{i \neq j}\big[E(X_i - \mu)E(X_j - \mu)\big]\Big] = \frac{\sigma^2}{n}.
\end{aligned}$$

$\Rightarrow$ 표본 평균 $\bar{X}$은 샘플 숫자가 커질수록 모집단의 평균 μ에 가까워진다.

표본평균의 성질

In [2]:
```
mu, sigma=0,1
```

In [3]:
```
np.random.normal(loc=mu, scale=sigma)
```

Out[3]: 0.5331692716674414

In [4]:
```
X=np.random.normal(loc=mu, scale=sigma, size=10000)
```

In [5]:
```
X #data
```

Out[5]:
```
array([ 0.71269468, -1.23367117,  1.22896926, ...,  2.10196696,
       -1.01883226,  1.13204636])
```

$\bar{X}$

In [6]:
```
X.mean()
```

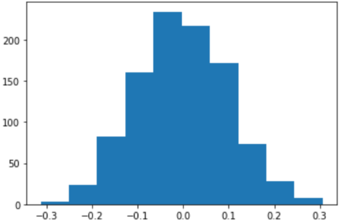
Out[6]: 0.0024860210670137267

표본평균의 성질

In [7]:
```
means=[np.mean(np.random.normal(loc=mu, scale=sigma,size=100)) for i in range(1000)]
```

In [8]:
```
plt.hist(means)
```

Out[8]:
```
(array([  3.,  24.,  82., 160., 234., 217., 171.,  73.,  28.,   8.]),
 array([-0.31337143, -0.25135704, -0.18934264, -0.12732824, -0.06531385,
        -0.00329945,  0.05871495,  0.12072934,  0.18274374,  0.24475814,
         0.30677254]),
 <BarContainer object of 10 artists>)
```



$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{N})$

In [9]:
```
np.mean(means)
```

Out[9]:
```
-0.002006202496045791
```

In [10]:
```
np.var(means) # close to 0.01 = 1/100
```

Out[10]:
```
0.010147261232552805
```

선형회귀분석

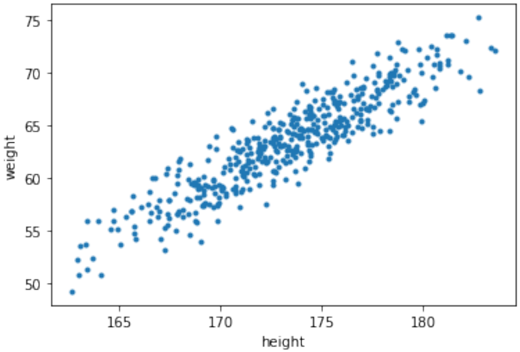
회귀분석

- 두개 이상의 변수 사이의 관계를 추정하기 위한 통계적 과정.

Process

- 종속변수 (dependent variable) 와 설명변수 (explanatory variable) 간의 수학적 모형 가정하기.
- 위에서 가정한 수학적 모형을 바탕으로 변수의 관측값(=데이터)을 이용하여 관계 추정하기.
- 추정된 모형이 데이터에 잘 맞는지 테스트 하기. 가정한 수학적 모형이 맞는지 확인하기.

회귀분석

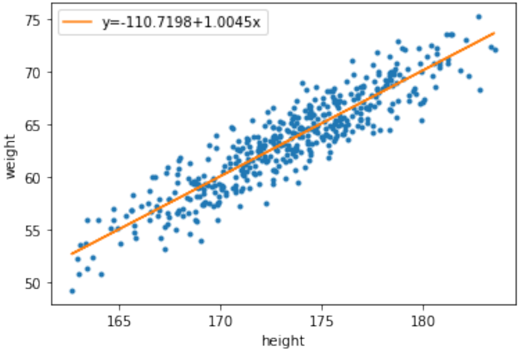


example

$$Y \text{ (종속 변수)} : \text{몸무게}$$
$$X \text{ (설명 변수)} : \text{키}$$

Relationship: $Y = f(X)$.

회귀분석



example

$$Y \text{ (종속 변수) : 몸무게}$$
$$X \text{ (설명 변수) : 키}$$

Relationship: $Y = f(X)$.　　선형 관계: $Y = \beta_0 + \beta_1 X$.

선형회귀분석

- 단순 선형 회귀분석: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ (β_0: intercept, β_1: slope)
  $E(y_i|X = x_i) = \beta_0 + \beta_1 x_i$
- 다중 선형 회귀분석: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i$ (β_0: intercept, $\beta_1, \cdots, \beta_p$: slopes)
  $E(y_i|X_1 = x_{1i}, X_2 = x_{2i}, \cdots, X_p = x_{pi}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}$

계수 $\beta_0, \beta_1, \beta_j$의 값을 추정해야 함.

$\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_N$은 모두 서로 독립이며 $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2$라 가정함.

단순 선형 회귀분석: 추정

데이터로부터 β_0와 β_1을 추정하는 방법? $\Rightarrow$ 최소제곱법 (Least Squares Method)

- 최소제곱법

$$\text{Find } \beta_0 \text{ and } \beta_1 \text{ that minimize } \sum_{i=1}^{N} \left\{ y_i - (\beta_0 + \beta_1 x_i) \right\}^2$$

$$\begin{aligned}
\sum_{i=1}^{N} \left\{ y_i - (\beta_0 + \beta_1 x_i) \right\}^2 &= \sum_{i=1}^{N} \left\{ y_i^2 - 2y_i(\beta_0 + \beta_1 x_i) + (\beta_0 + \beta_1 x_i)^2 \right\} \\
&= N\beta_0^2 - 2\beta_0 \sum_{i=1}^{N} (y_i - \beta_1 x_i) + \sum_{i=1}^{N} \{ y_i^2 - 2\beta_1 x_i y_i + \beta_1^2 x_i^2 \} \\
&= N\left\{ \beta_0^2 - 2\beta_0 \frac{1}{N} \sum_{i=1}^{N} (y_i - \beta_1 x_i) \right\} + \sum_{i=1}^{N} (y_i - \beta_1 x_i)^2 \\
&= N\left\{ \color{blue}{\beta_0} - \frac{1}{N} \sum_{i=1}^{N} (y_i - \beta_1 x_i) \right\}^2 - \frac{1}{N} \left\{ \sum_{i=1}^{N} (y_i - \beta_1 x_i) \right\}^2 + \sum_{i=1}^{N} (y_i - \beta_1 x_i
\end{aligned}$$

단순 선형 회귀분석: 추정

Set

$$\hat{\beta}_0 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \beta_1 x_i) = \bar{y} - \beta_1 \bar{x}.$$

이를 대입하면 아래와 같음.

$$R = -\frac{1}{N}\left\{\sum_{i=1}^{N}(y_i - \beta_1 x_i)\right\}^2 + \sum_{i=1}^{N}(y_i - \beta_1 x_i)^2.$$

$$\begin{aligned}
\frac{R}{N} &= -\left\{\frac{1}{N}\sum_{i=1}^{N} y_i - \beta_1 \frac{1}{N}\sum_{i=1}^{N} x_i\right\}^2 + \frac{1}{N}\sum_{i=1}^{N}(y_i - \beta_1 x_i)^2 \\
&= -\left\{\bar{y} - \beta_1 \bar{x}\right\}^2 + \frac{1}{N}\sum_{i=1}^{N}(y_i - \beta_1 x_i)^2 \\
&= -(\bar{y})^2 + 2\beta_1 \bar{x}\bar{y} - \beta_1^2(\bar{x})^2 + \frac{1}{N}\sum_{i=1}^{N} y_i^2 - \frac{2}{N}\beta_1 \sum_{i=1}^{N} x_i y_i + \beta_1^2 \frac{1}{N}\sum_{i=1}^{N} x_i^2 \\
&= \beta_1^2 \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 - 2\beta_1 \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) + C
\end{aligned}$$

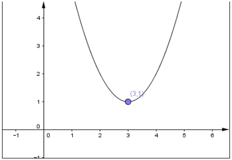
단순 선형 회귀분석: 추정

최소제곱 추정치:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} = \frac{S_{XY}}{S_X^2} = \frac{S_{XY}}{S_{XX}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

단순 선형 회귀분석: 추정



다른 계산방법: 미분 방정식

손실함수 $\sum_{i=1}^{N}\{y_i - (\beta_0 + \beta_1 x_i)\}^2$의 기울기를 0으로 두는 미분방정식을 세운다.

$$\frac{\partial}{\partial \beta_0}\sum_{i=1}^{N}\left\{y_i - (\beta_0 + \beta_1 x_i)\right\}^2 = -2\sum_{i=1}^{N}\left\{y_i - (\beta_0 + \beta_1 x_i)\right\} = 0$$

$$\frac{\partial}{\partial \beta_1}\sum_{i=1}^{N}\left\{y_i - (\beta_0 + \beta_1 x_i)\right\}^2 = -2\sum_{i=1}^{N} x_i\left\{y_i - (\beta_0 + \beta_1 x_i)\right\} = 0$$

즉, 아래의 연립방정식을 푼다.

$$\begin{cases} N\beta_0 + \sum_{i=1}^{N} x_i \beta_1 = \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i \beta_0 + \sum_{i=1}^{N} x_i^2 \beta_1 = \sum_{i=1}^{N} x_i y_i \end{cases}$$

단순 선형 회귀분석: 추정

행렬로 접근해보자

Let $Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \in \mathbb{R}^{N\times 1}$, $X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} \in \mathbb{R}^{N\times 2}$, $E = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix} \in \mathbb{R}^{N\times 1}$, and $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \in \mathbb{R}^{2\times 1}$.

$$\Rightarrow \quad Y = X\beta + E,$$

$$\begin{aligned}
\sum_{i=1}^{N} \left\{ y_i - (\beta_0 + \beta_1 x_i) \right\}^2 &= \left(Y - X\beta\right)^T \left(Y - X\beta\right) = \left(Y^T - (X\beta)^T\right)\left(Y - X\beta\right) \\
&= \left(Y^T - \beta^T X^T\right)\left(Y - X\beta\right) \quad (\because (AB)^T = B^T A^T) \\
&= Y^T Y - \beta^T X^T Y - Y^T X\beta + \beta^T X^T X\beta \\
&= Y^T Y - 2Y^T X\beta + \beta^T X^T X\beta
\end{aligned}$$

단순 선형 회귀분석: 추정

- 손실함수:

$$L(\beta) = Y^T Y - 2Y^T X\beta + \beta^T X^T X\beta$$

- 손실함수의 편미분:

$$\frac{\partial}{\partial \beta} L(\beta) = \begin{bmatrix} \frac{\partial}{\partial \beta_0} L(\beta) \\ \frac{\partial}{\partial \beta_1} L(\beta) \end{bmatrix} = -2X^T Y + 2X^T X\beta$$

- 최소제곱 추정치:

$$X^T X\beta = X^T Y\text{의 해:}$$

$$(X^T X)^{-1} X^T X\beta = (X^T X)^{-1} X^T Y$$

$$I\beta = (X^T X)^{-1} X^T Y$$

$$\Rightarrow \hat{\beta} = (X^T X)^{-1} X^T Y$$

기울기 강하(Gradient Descent)

기울기 강하 방법: 선형회귀분석에서는 손실함수를 최소화하는 β 값을 한번에 찾을 수 있지만 딥러닝에서는 그러지 못 함. 이때, 최적의 계수 β값을 찾을 때 손실함수의 기울기의 반대방향으로 β값을 조금씩 이동하면서 찾아나감. (즉, 기울기 강하 방법은 반복적인 과정을 통해 추정치를 도출함.)

- 손실함수:

$$L(\beta) = Y^T Y - 2Y^T X\beta + \beta^T X^T X\beta$$

- 손실함수의 편미분:

$$\frac{\partial}{\partial\beta}L(\beta) = \begin{bmatrix} \frac{\partial}{\partial\beta_0}L(\beta) \\ \frac{\partial}{\partial\beta_1}L(\beta) \end{bmatrix} = -2X^T Y + 2X^T X\beta$$

- 기울기 강하 (Gradient Descent) : 편미분(기울기) 반대방향으로의 이동

$$\begin{aligned} \hat{\beta}_{t+1} &= \hat{\beta}_t - \gamma \left.\frac{\partial}{\partial\beta}L(\beta)\right|_{\beta=\hat{\beta}_t} \\ &= \hat{\beta}_t - \gamma\left\{-2X^T Y + 2X^T X\hat{\beta}_t\right\} \end{aligned}$$

($\gamma > 0$ 를 학습률이라 함.)

기울기 강하(Gradient Descent)

- $f(x) \in \mathbb{R}$가 벡터 $x \in \mathbb{R}^d$ 의 함수라 하자.
- 함수의 편미분은 다음과 같다.

$$\frac{\partial}{\partial x_1} f(x), \frac{\partial}{\partial x_2} f(x), \cdots, \frac{\partial}{\partial x_d} f(x)$$

- **(임의의 유닛 벡터 u에 대한) 방향도함수(Directional derivative)** $= u$의 방향으로 변화하는 정도

$$\begin{aligned} D_u f(a) &= \left[\frac{\partial}{\partial x} f(x) \Big|_{x=a} \right]^T u \\ &= \left|\left| \frac{\partial}{\partial x} f(x) \Big|_{x=a} \right|\right| ||u|| cos\theta \end{aligned}$$

- $D_u f(a)$ 는 $\theta = 0^\circ$ 일때 가장 크고 $\theta = 180^\circ$ 일때 가장 작음.

다중 선형 회귀분석: 추정

Let $Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \in \mathbb{R}^{N\times 1}$, $X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1N} & x_{2N} & \cdots & x_{pN} \end{bmatrix} \in \mathbb{R}^{N\times(p+1)}$,

$E = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix} \in \mathbb{R}^{N\times 1}$, and $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \in \mathbb{R}^{(p+1)\times 1}$.

$$\Rightarrow \quad Y = X\beta + E,$$

$$\begin{aligned} \sum_{i=1}^{N} \{y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi})\}^2 &= (Y - X\beta)^T (Y - X\beta) \\ &= Y^T Y - 2Y^T X\beta + \beta^T X^T X\beta \end{aligned}$$

다중 선형 회귀분석: 추정

최소제곱추정치는 $X^T X\beta = X^T Y$의 해다.

$$X^T X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pN} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1N} & x_{2N} & \cdots & x_{pN} \end{bmatrix}$$

$$= \begin{bmatrix} N & \sum_i x_{1i} & \sum_i x_{2i} & \cdots & \sum_i x_{pi} \\ \sum_i x_{1i} & \sum_i x_{1i}^2 & \sum_i x_{1i}x_{2i} & \cdots & \sum_i x_{1i}x_{pi} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_i x_{pi} & \sum_i x_{1i}x_{pi} & \sum_i x_{2i}x_{pi} & \cdots & \sum_i x_{pi}^2 \end{bmatrix} \in \mathbb{R}^{(p+1)\times(p+1)}$$

$$X^T Y = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pN} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} \sum_i y_i \\ \sum_i x_{1i}y_i \\ \sum_i x_{2i}y_i \\ \vdots \\ \sum_i x_{pi}y_i \end{bmatrix} \in \mathbb{R}^{(p+1)\times 1}$$

다중 선형 회귀분석: 추정

$X^TX\beta = X^TY \Leftrightarrow$:

$$
\begin{array}{ccccccccccc}
\beta_0 N & + & \beta_1 \sum_i x_{1i} & + & \beta_2 \sum_i x_{2i} & + & \cdots & + & \beta_p \sum_i x_{pi} & = & \sum_i y_i \\
\beta_0 \sum_i x_{1i} & + & \beta_1 \sum_i x_{1i}^2 & + & \beta_2 \sum_i x_{1i}x_{2i} & + & \cdots & + & \beta_p \sum_i x_{1i}x_{pi} & = & \sum_i x_{1i}y_i \\
\vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\
\beta_0 \sum_i x_{pi} & + & \beta_1 \sum_i x_{1i}x_{pi} & + & \beta_2 \sum_i x_{2i}x_{pi} & + & \cdots & + & \beta_p \sum_i x_{pi}^2 & = & \sum_i x_{pi}y_i
\end{array}
$$

만약 $(X^TX)^{-1}$가 존재하면,

$$\hat{\beta} = (X^TX)^{-1}X^TY.$$

기울기 강하(Gradient Descent)

- 손실함수:

$$L(\beta) = \sum_{i=1}^{N} \{y_i - (\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})\}^2 = Y^T Y - 2Y^T X\beta + \beta^T X^T X\beta$$

- 손실함수의 편미분:

$$\frac{\partial}{\partial\beta} L(\beta) = \begin{bmatrix} \frac{\partial}{\partial\beta_0} L(\beta) \\ \frac{\partial}{\partial\beta_1} L(\beta) \\ \vdots \\ \frac{\partial}{\partial\beta_p} L(\beta) \end{bmatrix} = \begin{bmatrix} -2\sum_{i=1}^{N} \{y_i - (\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_p)\} \\ -2x_{1i}\sum_{i=1}^{N} \{y_i - (\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_p)\} \\ \vdots \\ -2x_{pi}\sum_{i=1}^{N} \{y_i - (\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_p)\} \end{bmatrix}$$

$$\frac{\partial}{\partial\beta} L(\beta) = -2X^T Y + 2X^T X\beta = -2X^T(Y - X\beta)$$

- 기울기 강하 (Gradient Descent) : 편미분(기울기) 반대방향으로의 이동

최대가능도 추정

- 손실함수:

$$L(\beta) = \sum_{i=1}^{N} \{y_i - (\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})\}^2 = Y^T Y - 2Y^T X\beta + \beta^T X^T X\beta$$

- 위의 식을 손실함수로 쓰는 근거? ⇒ 최대가능도 (Maximum Likelihood) 추정치를 도출하기 때문에!

최대가능도 (Maximum Likelihood) 방법

- 가능도 함수(Likelihood): 데이터의 확률분포를 모수(β)의 함수로 표현한 것.

  예) 선형회귀분석 문제를 다시 생각해보자.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$

  여기서 확률 변수 ε_i가 독립이며, 정규분포 $\mathcal{N}(0, \sigma^2)$을 따른다고 가정하자.
  즉, $y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}, \sigma^2)$ 이며 확률밀도함수는

$$p(y_i) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{1}{2}\left(\frac{y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_p x_{pi}}{\sigma}\right)^2}$$

  또, $y_1, y_2, \cdots, y_N$의 결합 확률밀도함수는

$$p(y_1, y_2, \cdots, y_N) = \prod_{i=1}^{N} p(y_i) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{1}{2}\left(\frac{y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_p x_{pi}}{\sigma}\right)^2}$$

$$\Rightarrow Likelihood(\beta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{1}{2}\left(\frac{y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_p x_{pi}}{\sigma}\right)^2}$$

최대가능도 (Maximum Likelihood) 추정치

- 최대가능도 추정치 (Maximum Likelihood Estimate; MLE): 가능도 함수를 최대로 하는 모수의 값.

  예) 선형회귀분석 문제의 MLE 구하기

  $Likelihood(\beta)$ 를 최대로 하는 β 값과 $\log Likelihood(\beta)$ 를 최대로 하는 β 값은 동일하므로, $\log Likelihood(\beta)$를 구해보면 아래와 같다:

$$\log Likelihood(\beta) = \sum_{i=1}^{N} \left\{ \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2} \left(\frac{y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_p x_{pi}}{\sigma} \right)^2 \right\}$$

  $\log Likelihood(\beta)$를 최대로 하는 β 값은 $L(\beta) = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_p x_{pi})^2$를 가장 작게 하는 최소제곱추정치와 동일함을 알 수 있다!

최대가능도 (Maximum Likelihood) 추정치

- 최대가능도 추정치 (Maximum Likelihood Estimate) 를 사용하는 이유
  1. 데이터 개수 N이 커지면 추정치가 실제값으로 확률 수렴(consistent)한다.
  2. 점근적으로 효율적인 (asymptotically efficient) 추정치다.

로지스틱 회귀분석

로지스틱 회귀분석

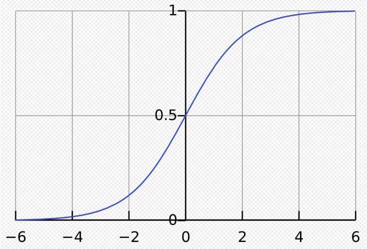
- 로지스틱 회귀분석 문제:

$$y_i \sim Ber\left(\pi(x_i)\right), \quad \pi(x_i) = \frac{\exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi})}$$

즉, y_i는 0과 1중 하나의 값을 가지는 이항 확률 변수이며

$$P(y_i = 1|x_i) = \pi(x_i), \quad P(y_i = 0|x_i) = 1 - \pi(x_i)$$

이다.



로지스틱 회귀분석

- 이때 y_i의 확률 질량 함수는

$$p(y_i|x_i) = \pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i}$$

이며 $y_1, y_2, \cdots, y_N$ 들이 독립이라고 가정하면 결합 확률 질량 함수는

$$p(y_1, \cdots, y_N|x_1, \cdots, x_N) = \prod_{i=1}^{N} p(y_i|x_i) = \prod_{i=1}^{N} \pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i}$$

$$\Rightarrow Likelihood(\beta) = \prod_{i=1}^{N} \pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i}$$

$$\Rightarrow \log Likelihood(\beta) = \sum_{i=1}^{N} \{y_i \log\pi(x_i) + (1-y_i)\log(1-\pi(x_i))\}$$

로지스틱 회귀분석

- 기울기 강하 등의 방법을 통해 $\log Likelihood(\beta)$를 최대로 하는 β의 값을 찾을 수 있다.
- -$\log Likelihood(\beta)$ 를 교차 엔트로피 (cross-entropy) 손실함수라고 부르기도 한다.
- $-\log Likelihood(\beta)$의 편미분은?

퀴즈!

로지스틱 회귀분석

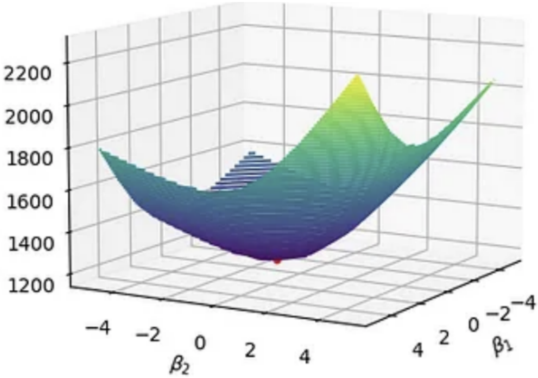


Figure: Graph of Negative log likelihood (ref: Nicolo Cosimo Albanese, "Negative log-partial likelihood of the Cox model with local optimum", towardsdatascience, Published date: Dec 10, 2022, https://towardsdatascience.com/survival-analysis-optimize-the-partial-likelihood-of-the-cox-model-b56b8f112401)