선형회귀 (Linear Regression) 모델 - 2

2025 Spring

머신러닝1

이 두 호

개요

- 선형회귀 모델 1
  - 선형회귀 모델 개요
  - 선형회귀 모델 가정
- 선형회귀 모델 2 ✓
  - 파라미터 추정 : 최소제곱법
  - 파라미터 추정 : 경사하강법
- 선형회귀 모델 3
  - 결정계수 (R^2)
  - 분산분석

$$Y = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_p x_p + \epsilon$$

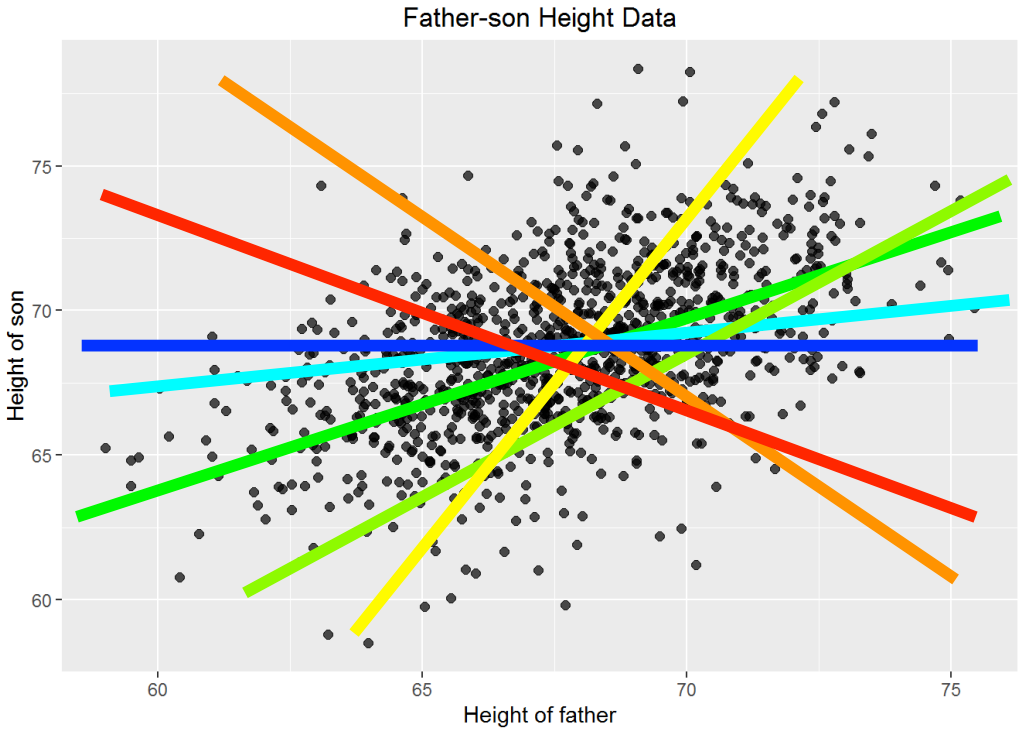
$$E[Y] = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_p x_p$$

입력변수(X)와 출력변수(Y) 평균(기대값)

사이의 관계를 정량화하여 선형식으로 표현하기!

그래서 선형식은 뭔디?

$$E[Y] = f(X) = w_0 + w_1 X$$

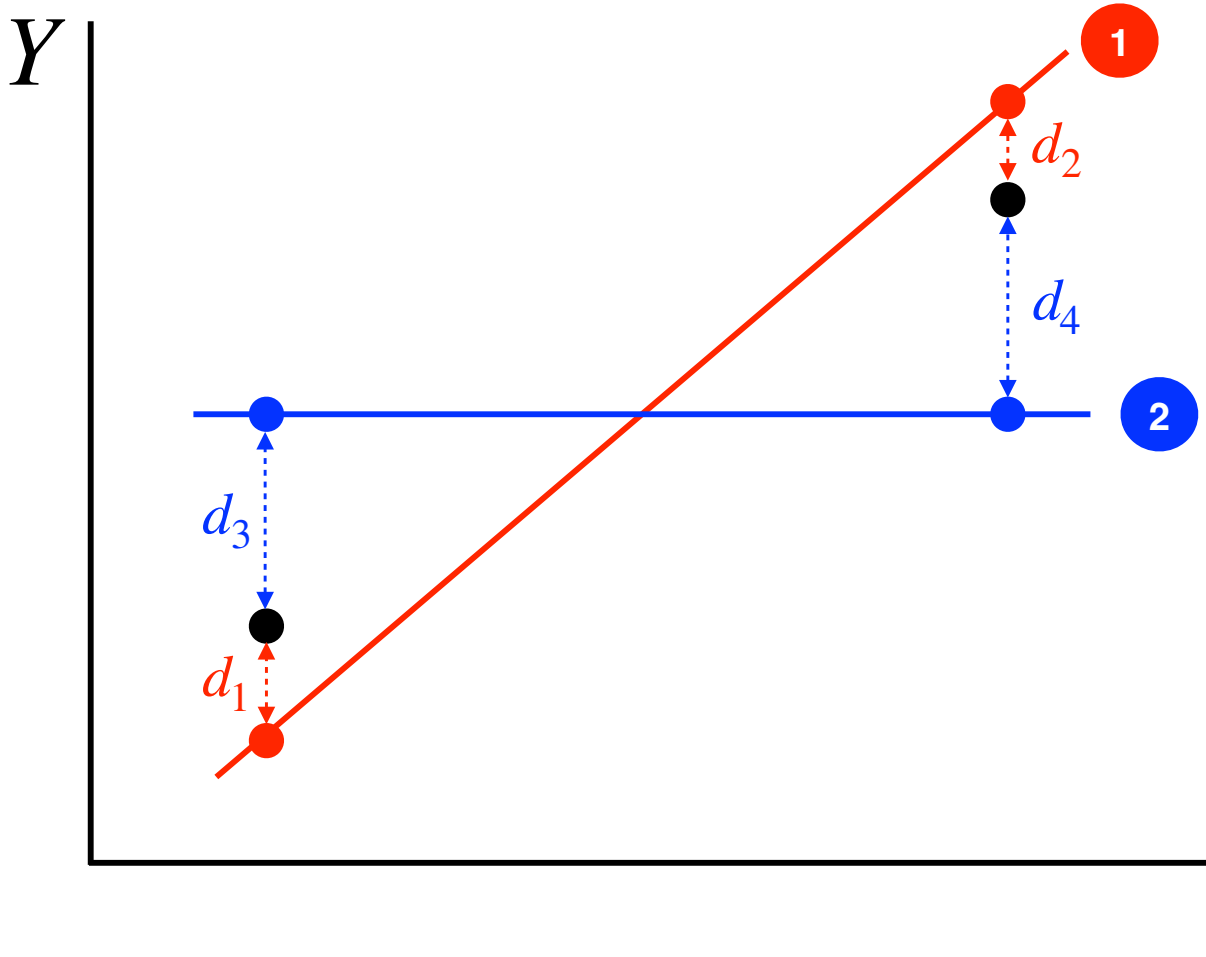


파라미터 (parameter)

파라미터를 찾자 (추정, estimation)

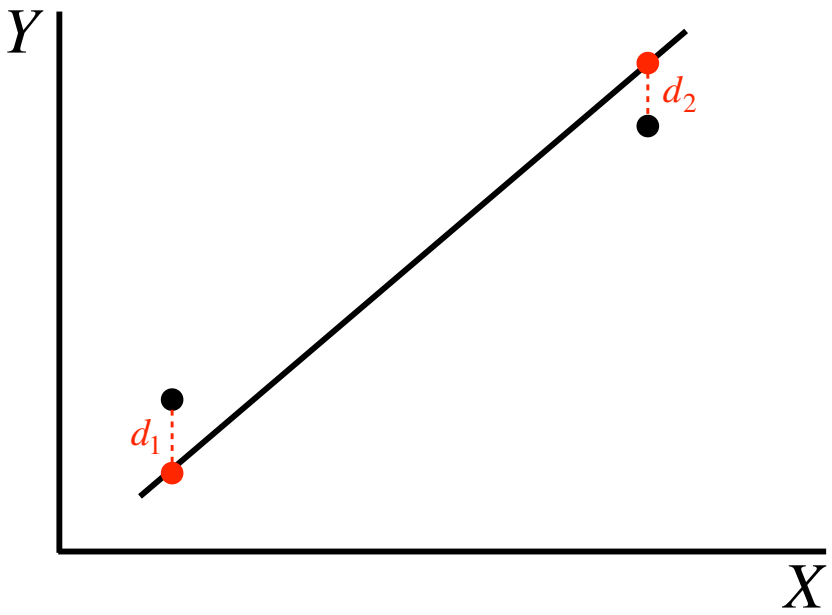
우리가 가지고 있는 데이터들의 함수식으로!

파라미터 추정



$$d_1 + d_2 < d_3 + d_4$$

파라미터 추정



$$d_1 + d_2 + \cdots + d_n = 0$$

$$d_1^2 + d_2^2 + \cdots + d_n^2 \geq 0$$

$$d_i = Y_i - E[Y_i] = Y_i - (w_0 + w_1 X_i)$$

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} \{Y_i - (w_0 + w_1 X_i)\}^2$$

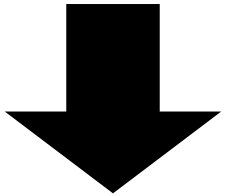
$$\min_{w_0, w_1 \in \mathbb{R}} \underbrace{\sum_{i=1}^{n} \{Y_i - (w_0 + w_1 X_i)\}^2}$$

Cost function (비용함수)

파라미터 추정

Cost function (비용함수)

$$\min_{w_0,\, w_1 \in \mathbb{R}} \sum_{i=1}^{n} \left\{ Y_i - (w_0 + w_1 X_i) \right\}^2$$



Algorithms

$$\hat{w}_0,\ \hat{w}_1$$

$$\hat{f}(X) = \hat{w}_0 + \hat{w}_1 X$$

파라미터 추정 알고리즘 (최소제곱법, Least square estimation, LSE)

Cost function (비용함수)

$$\min_{w_0, w_1 \in \mathbb{R}} \sum_{i=1}^{n} \left\{ Y_i - (w_0 + w_1 X_i) \right\}^2$$

Cost function is **Convex** → Globally optimal solution exists.

Let $C(w_0, w_1) = \sum_{i=1}^{n} \left\{ Y_i - (w_0 + w_1 X_i) \right\}^2$. Then, we have

$$\frac{\partial C(w_0, w_1)}{\partial w_0} = -2 \sum_{i=1}^{n} \left\{ Y_i - (w_0 + w_1 X_i) \right\}$$

$$\frac{\partial C(w_0, w_1)}{\partial w_1} = -2 \sum_{i=1}^{n} \left\{ Y_i - X_i(w_0 + w_1 X_i) \right\}$$

파라미터 추정 알고리즘 (최소제곱법, Least square estimation, LSE)

$$\frac{\partial C(w_0, w_1)}{\partial w_0} = -2\sum_{i=1}^{n}\left\{Y_i - (w_0 + w_1 X_i)\right\} = 0$$

$$\frac{\partial C(w_0, w_1)}{\partial w_1} = -2\sum_{i=1}^{n}\left\{Y_i - X_i(w_0 + w_1 X_i)\right\} = 0$$

$$\hat{w}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}, \quad \hat{w}_0 = \overline{Y} - \hat{w}_1\overline{X}$$

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}, \quad \overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

$$f(X) = E[Y] = \hat{Y} = \hat{w}_0 + \hat{w}_1 X$$

- LSE Algorithm

Step 1. Cost function: $C(w_0, w_1) = \sum_{i=1}^{n} \{Y_i - (w_0 + w_1 X_i)\}^2$.

Step 2. Optimization problem: $\min_{w_0, w_1 \in \mathbb{R}} C(w_0, w_1)$.

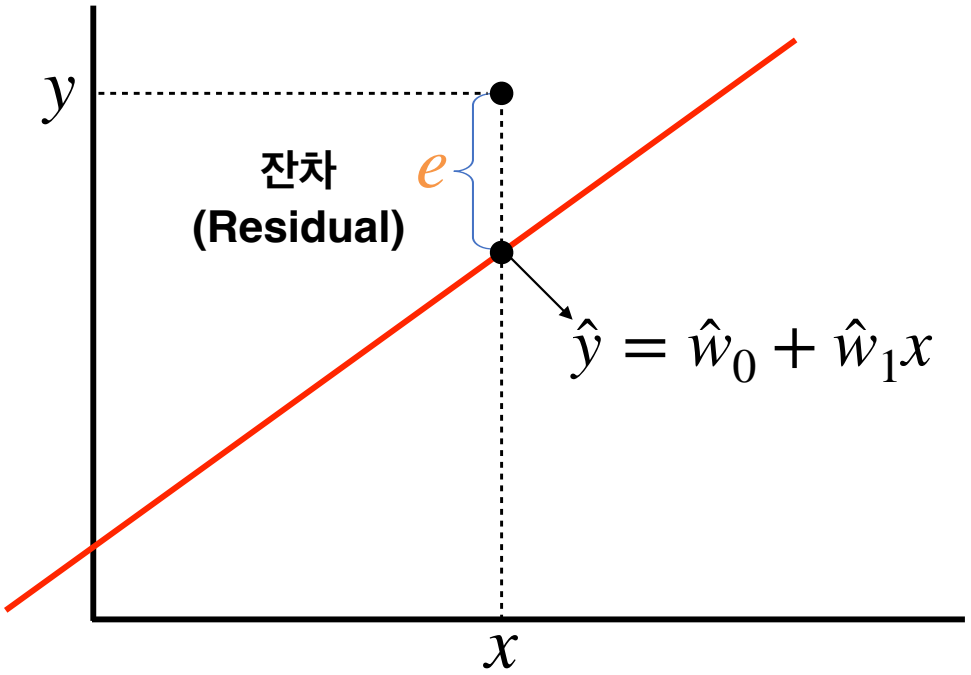
Step 3. Solving equations:

$$\frac{\partial C(w_0, w_1)}{\partial w_0} = -2 \sum_{i=1}^{n} \{Y_i - (w_0 + w_1 X_i)\} = 0$$

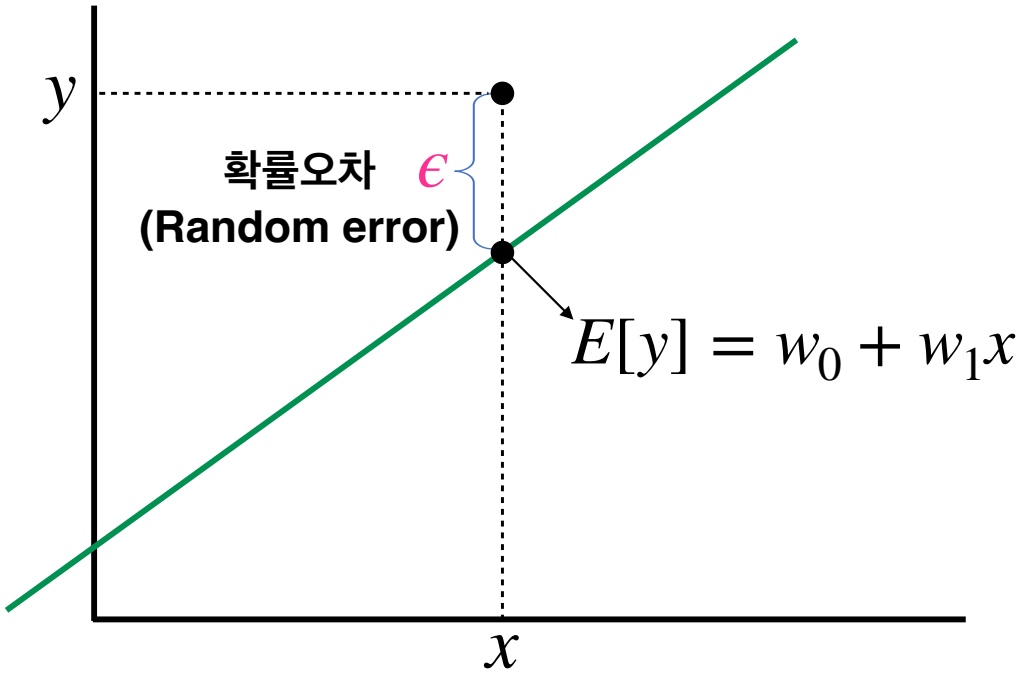
$$\frac{\partial C(w_0, w_1)}{\partial w_1} = -2 \sum_{i=1}^{n} \{Y_i - X_i(w_0 + w_1 X_i)\} = 0$$

Solutions: $\hat{w}_1 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$, $\hat{w}_0 = \overline{Y} - \hat{w}_1 \overline{X}$

잔차 (Residual)



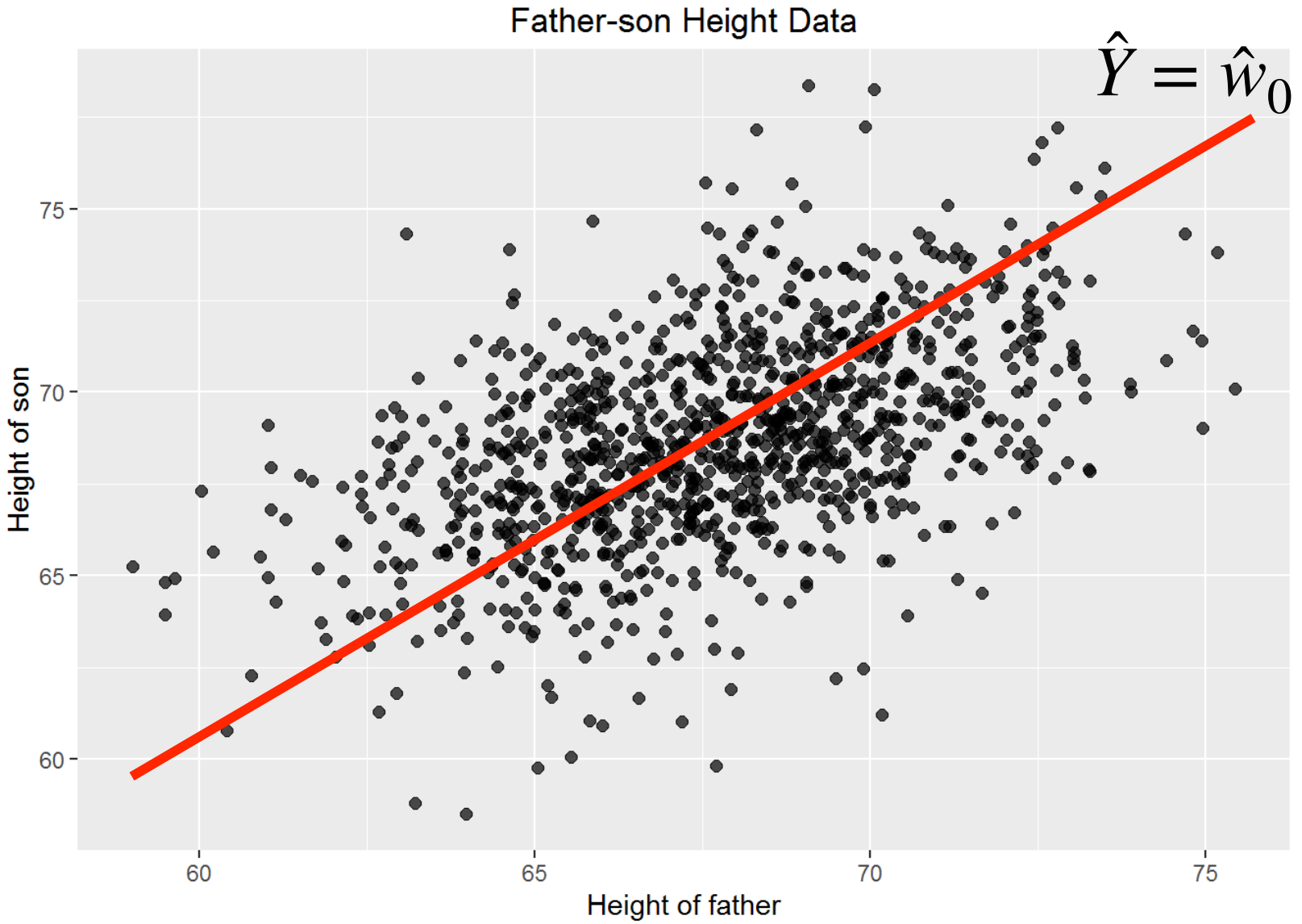
$e = y - \hat{y}$



$\epsilon = y - E[y]$

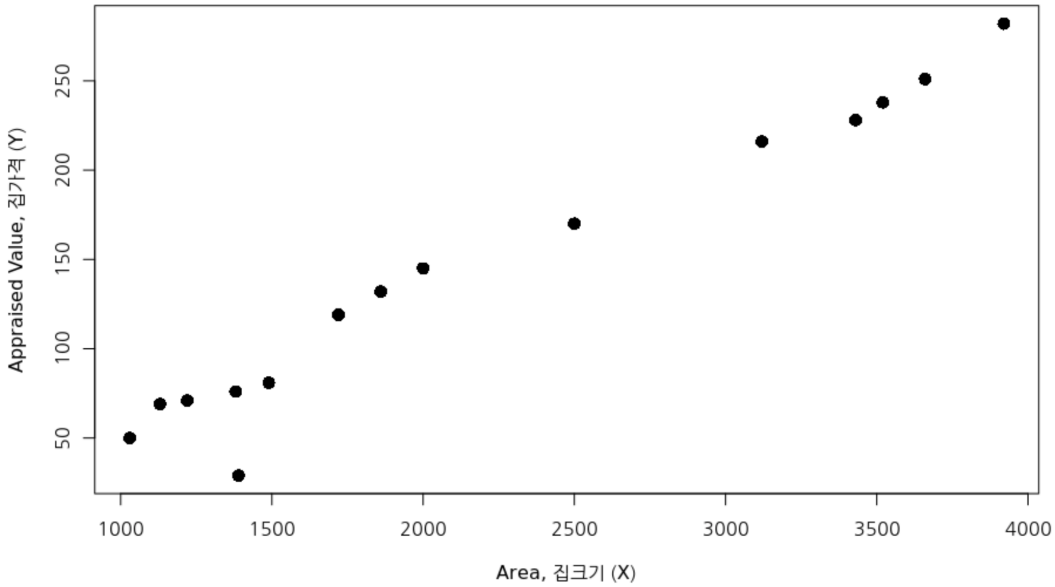
잔차 e 는 확률오차 ϵ 이 실제로 구현된 값이다!

학습 결과 직선


Father-son Height Data
$\hat{Y} = \hat{w}_0 + \hat{w}_1 X$
Height of son
75
70
65
60
60
65
70
75
Height of father

선형회귀 모델 예제

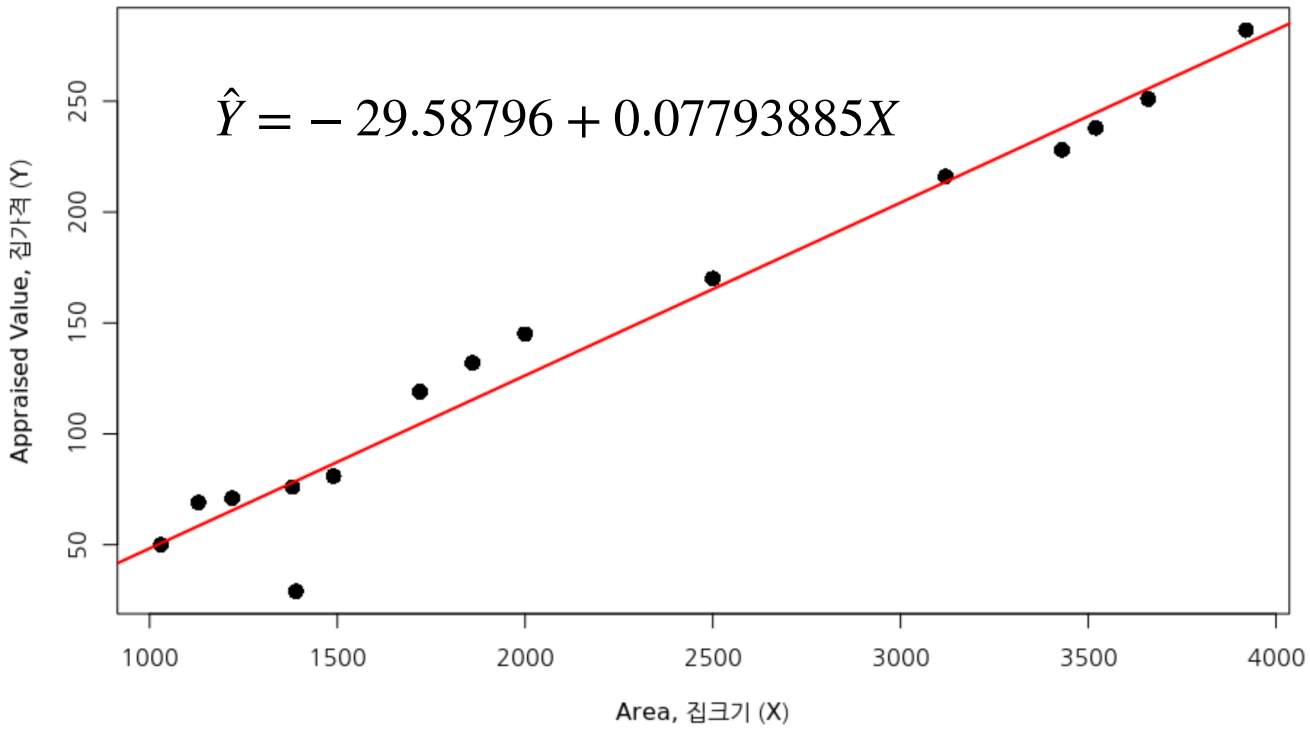
관측치	Area, 집크기 (X)	Appraised Value, 집가격 (Y)
1	1380	76
2	3120	216
3	3520	238
4	1130	69
5	1030	50
6	1720	119
7	3920	282
8	1490	81
9	1860	132
10	3430	228
11	2000	145
12	3660	251
13	2500	170
14	1220	71
15	1390	29



$$\hat{w}_1 = 0.07793885, \quad \hat{w}_0 = -29.58796$$

집 크기가 1 m² 증가할 때 마다 집 가격은 0.07794 증가

선형회귀 모델 예제



1. 집 크기가 2227 m² 일 때, 집가격을 예측하라.

2. 현재 150 억이 있으면 얼마나 큰 집을 살 수 있는가?

3. 각 관측에서의 잔차를 구하고, 잔차들의 합을 구하라.

HW

x	1	2	3	4	5	6	7	8	9	10	11
y	3	3	3	6	6	9	9	9	10	11	?

회귀분석을 이용하여 x 값이 11 일 때, y 값을 예측하라.

- 10개 기업을 대상으로 한 해 동안의 광고비와 신규고객 수를 조사

기업	X: 광고비(억 원)	Y: 신규고객(100명)
A	36.5	14
B	28.0	9
C	42.9	15
D	52.0	20
E	51.5	21
F	53.8	25
G	25.4	9
H	37.2	13
I	50.9	20
J	29.2	10

신규고객 1,700 명을 확보하기 위해 얼마의 광고비를 지출해야 하나?

입력변수가 2개 이상이면? 다중선형회귀분석

i	X_1	X_2	...	X_p	Y
1	X_{11}	X_{12}	...	X_{1p}	Y_1
2	X_{21}	X_{22}	...	X_{2p}	Y_2
⋮	⋮	⋮	⋱	⋮	⋮
n	X_{n1}	X_{n2}	...	X_{np}	Y_n

$$Y_i = w_0 + w_1 x_{1i} + w_2 x_{2i} + \cdots + w_p x_{pi} + \epsilon_i, \quad \forall i$$

Cost function (비용함수)

$$C(w_0, w_1, \cdots, w_p) = \sum_{i=1}^{n} \left\{ Y_i - (w_0 + w_1 x_{1i} + w_2 x_{2i} + \cdots + w_p x_{pi}) \right\}^2$$

입력변수가 2개 이상이면? 다중선형회귀분석

Cost function (비용함수)

$$C(w_0, w_1, \cdots, w_p) = \sum_{i=1}^{n} \left\{ Y_i - (w_0 + w_1 x_{1i} + w_2 x_{2i} + \cdots + w_p x_{pi}) \right\}^2$$

$$\min_{w_0, \cdots, w_p \in \mathbb{R}} \quad C(w_0, w_1, \cdots, w_p)$$

Cost function is **Convex** → Globally optimal solution exists.

$$\frac{\partial C}{\partial w_0} = 0, \quad \frac{\partial C}{\partial w_1} = 0, \quad \cdots, \quad \frac{\partial C}{\partial w_p} = 0.$$

입력변수가 1개 이상이면? 다중선형회귀분석

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}_{n \times (1+p)} \quad \mathbf{y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n \times 1} \quad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_p \end{bmatrix}_{(1+p) \times 1}$$

$$C(w_0, w_1, \cdots, w_p) = C(\mathbf{w}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\mathbf{y} - \mathbf{X}\mathbf{w})$$

$$= \mathbf{y}^\top \mathbf{y} - \mathbf{w}^\top \mathbf{X}^\top \mathbf{y} - \mathbf{y}^\top \mathbf{X}\mathbf{w} + \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w}$$

$$= \mathbf{y}^\top \mathbf{y} - 2\mathbf{w}^\top \mathbf{X}^\top \mathbf{y} + \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w}$$

$$\frac{\partial C(\mathbf{w})}{\partial \mathbf{w}} = -2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\mathbf{w} = \mathbf{0}$$

입력변수가 1개 이상이면? 다중선형회귀분석

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}_{n \times (1+p)} \quad \mathbf{y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n \times 1} \quad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_p \end{bmatrix}_{(1+p) \times 1}$$

$$\frac{\partial C(\mathbf{w})}{\partial \mathbf{w}} = -2\mathbf{X}^{\top}\mathbf{y} + 2\mathbf{X}^{\top}\mathbf{X}\mathbf{w} = \mathbf{0} \quad \Rightarrow \quad \mathbf{X}^{\top}\mathbf{X}\mathbf{w} = \mathbf{X}^{\top}\mathbf{y}$$

$$\hat{\mathbf{w}} = (\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}\mathbf{y}, \text{ and } \hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{w}}$$

다중선형회귀분석 예제

	x1	x2	y
1	26	1010	56.38
2	5	745	20.64
3	8	809	28.24
4	2	451	8.59
5	3	427	14.23
6	24	308	45.84
7	12	1408	43.13
8	15	1263	49.15
9	19	1440	50.68
10	20	1356	59.12
11	27	725	54.42
12	11	678	26.61
13	23	537	45.23
14	22	499	46.20
15	28	515	55.81
16	10	853	30.30
17	6	406	15.03
18	14	858	35.73
19	18	266	39.40
20	25	1348	63.61
21	13	908	35.09
22	17	519	34.53
23	30	718	60.93
24	16	585	31.55
25	21	971	49.27

```
set.seed(1)
x1 <- sample(2:30, 25)
set.seed(2)
x2 <- sample(36:1460, 25)
set.seed(3)
y <- round(2.341 + 1.6159*x1 + 0.015*x2 + rnorm(25,0,3.25),2)
df1 <- data.frame(x1,x2,y)
View(df1)
```

1. 다중선형회귀분석을 실시하라.

$$\hat{\mathbf{w}} = (\hat{w}_0, \hat{w}_1, \hat{w}_2) = ?$$

2. 각 관측치에서 잔차를 구하고, 잔차들의 합을 구하라.

HW

state <- as.data.frame(state.x77)

	Population	Income	Illiteracy	Life Exp	Murder	HS Grad	Frost	Area
Alabama	3615	3624	2.1	69.05	15.1	41.3	20	50708
Alaska	365	6315	1.5	69.31	11.3	66.7	152	566432
Arizona	2212	4530	1.8	70.55	7.8	58.1	15	113417
Arkansas	2110	3378	1.9	70.66	10.1	39.9	65	51945
California	21198	5114	1.1	71.71	10.3	62.6	20	156361
Colorado	2541	4884	0.7	72.06	6.8	63.9	166	103766
Connecticut	3100	5348	1.1	72.48	3.1	56.0	139	4862
Delaware	579	4809	0.9	70.06	6.2	54.6	103	1982
Florida	8277	4815	1.3	70.66	10.7	52.6	11	54090
Georgia	4931	4091	2.0	68.54	13.9	40.6	60	58073
Hawaii	868	4963	1.9	73.60	6.2	61.9	0	6425
Idaho	813	4119	0.6	71.87	5.3	59.5	126	82677
Illinois	11197	5107	0.9	70.14	10.3	52.6	127	55748
Indiana	5313	4458	0.7	70.88	7.1	52.9	122	36097
Iowa	2861	4628	0.5	72.56	2.3	59.0	140	55941
Kansas	2280	4669	0.6	72.58	4.5	59.9	114	81787
Kentucky	3387	3712	1.6	70.10	10.6	38.5	95	39650
Louisiana	3806	3545	2.8	68.76	13.2	42.2	12	44930
Maine	1058	3694	0.7	70.39	2.7	54.7	161	30920
Maryland	4122	5299	0.9	70.22	8.5	52.3	101	9891
Massachusetts	5814	4755	1.1	71.83	3.3	58.5	103	7826
Michigan	9111	4751	0.9	70.63	11.1	52.8	125	56817
Minnesota	3921	4675	0.6	72.96	2.3	57.6	160	79289
Mississippi	2341	3098	2.4	68.09	12.5	41.0	50	47296

Population, Income, Illiteracy, Life Exp, Frost 를 입력변수로 하고, Murder 를 출력변수로 하여 경사하강법을 이용해 다중회귀분석을 실시하라.

경사하강법 응용 : 선형회귀모델

i	X_1	X_2	...	X_p	Y
1	X_{11}	X_{12}	...	X_{1p}	Y_1
2	X_{21}	X_{22}	...	X_{2p}	Y_2
⋮	⋮	⋮	⋱	⋮	⋮
n	X_{n1}	X_{n2}	...	X_{np}	Y_n

$$Y_i = w_0 + w_1 x_{1i} + w_2 x_{2i} + \cdots + w_p x_{pi} + \epsilon_i, \quad \forall i$$

Cost function (비용함수)

$$C(w_0, w_1, \cdots, w_p) = \frac{1}{2n} \sum_{i=1}^{n} \left\{ Y_i - (w_0 + w_1 x_{1i} + w_2 x_{2i} + \cdots + w_p x_{pi}) \right\}^2$$

경사하강법 응용 : 선형회귀모델

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}_{n \times (1+p)} \quad \mathbf{y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n \times 1} \quad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_p \end{bmatrix}_{(1+p) \times 1}$$

$$C(w_0, w_1, \cdots, w_p) = C(\mathbf{w}) = \frac{1}{2n}(\mathbf{y} - \mathbf{X}\mathbf{w})^\top(\mathbf{y} - \mathbf{X}\mathbf{w})$$

$$= \frac{1}{2n}\left(\mathbf{y}^\top\mathbf{y} - \mathbf{w}^\top\mathbf{X}^\top\mathbf{y} - \mathbf{y}^\top\mathbf{X}\mathbf{w} + \mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w}\right)$$

$$= \frac{1}{2n}\left(\mathbf{y}^\top\mathbf{y} - 2\mathbf{w}^\top\mathbf{X}^\top\mathbf{y} + \mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w}\right)$$

$$\frac{\partial C(\mathbf{w})}{\partial \mathbf{w}} = \frac{\mathbf{X}^\top\left(\mathbf{X}\mathbf{w} - \mathbf{y}\right)}{n}$$

경사하강법 응용 : 선형회귀모델

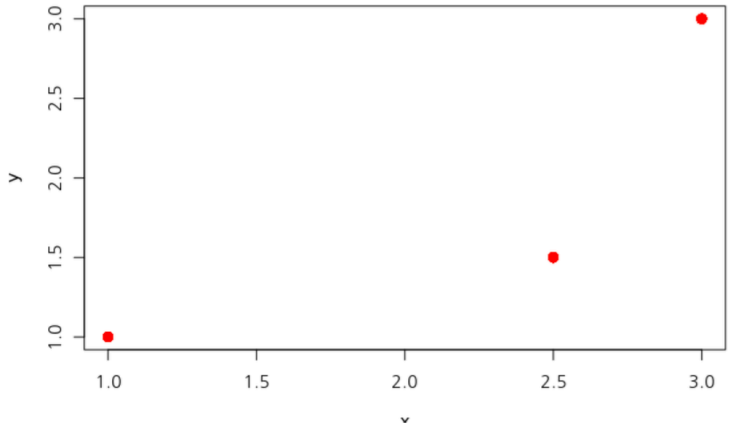
$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}_{n \times (1+p)} \quad \mathbf{y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n \times 1} \quad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_p \end{bmatrix}_{(1+p) \times 1}$$

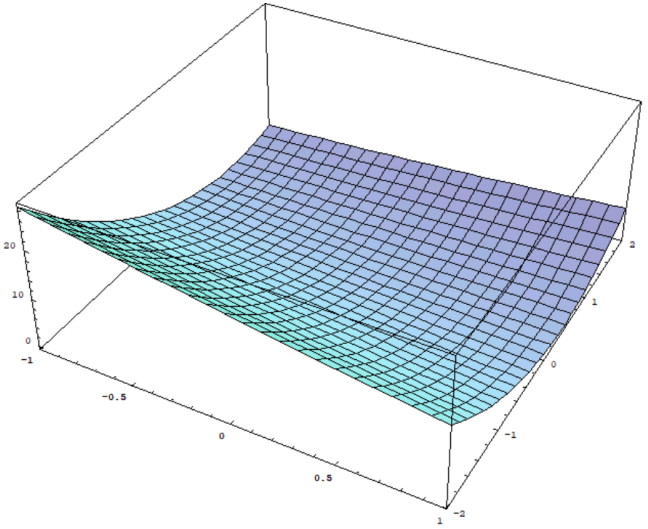
$$\frac{\partial C(\mathbf{w})}{\partial \mathbf{w}} = \frac{\mathbf{X}^\top \left(\mathbf{X}\mathbf{w} - \mathbf{y} \right)}{n}$$

$$\mathbf{w} \leftarrow \mathbf{w} - \alpha \frac{\partial C(\mathbf{w})}{\partial \mathbf{w}} \quad \Rightarrow \quad \mathbf{w} \leftarrow \mathbf{w} - \frac{\alpha \mathbf{X}^\top \left(\mathbf{X}\mathbf{w} - \mathbf{y} \right)}{n}$$

경사하강법 응용 : 선형회귀모델 예제

i	X_1	Y
1	1	1
2	2.5	1.5
3	3	3



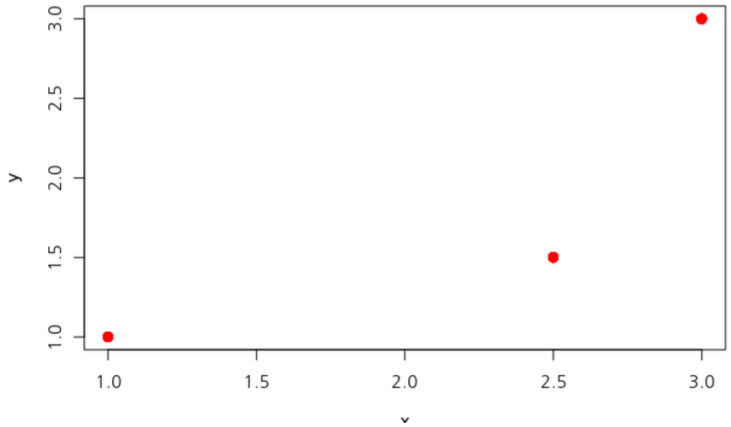


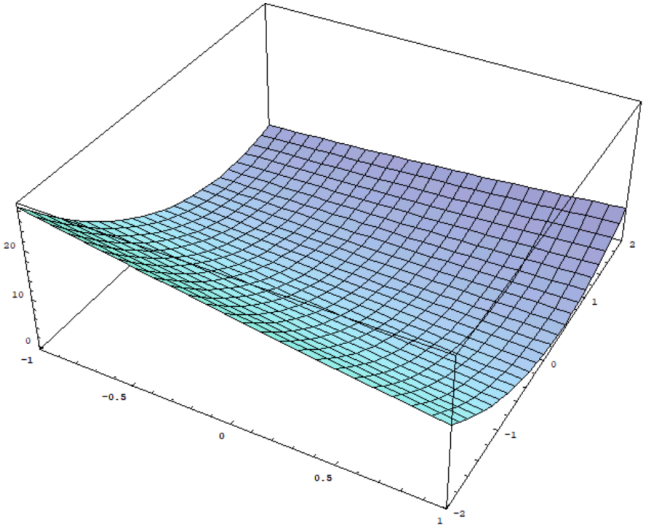
$$C(w_0, w_1) = \frac{1}{6}\left[\{1-(w_0+w_1\times 1)\}^2+\{1.5-(w_0+w_1\times 2.5)\}^2+\{3-(w_0+w_1\times 3)\}^2\right]$$

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 2.5 \\ 1 & 3 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 1 \\ 1.5 \\ 3 \end{bmatrix} \quad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

경사하강법 응용 : 선형회귀모델 예제

i	X_1	Y
1	1	1
2	2.5	1.5
3	3	3





$$\mathbf{w} \leftarrow \mathbf{w} - \frac{\alpha \mathbf{X}^\top (\mathbf{X}\mathbf{w} - \mathbf{y})}{n}$$

$$\mathbf{w}_0 = (0,0)^\top, \quad \alpha = 0.07$$

$$\mathbf{w}_1 = \mathbf{w}_0 - \frac{\alpha \mathbf{X}^\top (\mathbf{X}\mathbf{w}_0 - \mathbf{y})}{n} = (0.1283333, 0.3208333)$$

$$\mathbf{w}_2 = \mathbf{w}_1 - \frac{\alpha \mathbf{X}^\top (\mathbf{X}\mathbf{w}_1 - \mathbf{y})}{n} = (0.1990236, 0.5005535)$$

$$\vdots$$

$$\mathbf{w}_{1000} = \mathbf{w}_{999} - \frac{\alpha \mathbf{X}^\top (\mathbf{X}\mathbf{w}_{999} - \mathbf{y})}{n} = (9.413314 \times 10^{-5}, 0.8461154)$$

LSE algorithm

$$\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = (0, 0.8462)$$

$$\hat{\mathbf{w}} \approx \mathbf{w}_{1000}$$

경사하강법 알고리즘 종료조건

Step 1. x 값을 초기화 한다.

Step 2. x 값을 다음과 같이 갱신한다. : $x \leftarrow x - \alpha \cdot \frac{df}{dx}$

Step 3. **종료조건**에 해당되면 $x^* \leftarrow x$ 이고, 그렇지 않으면 Step 2 를 반복한다.

반복

1.반복횟수 지정하여 종료

2.$\mathbf{w}$ 의 변화가 굉장히 작으면 종료 : $\frac{| w_{i+1} - w_i |}{w_i} \ll \epsilon, \ \forall i$

3.$C(\mathbf{w})$ 의 변화가 굉장히 작으면 종료 : $\frac{| C(\mathbf{w}_{i+1}) - C(\mathbf{w}_i) |}{C(\mathbf{w}_i)} \ll \epsilon$