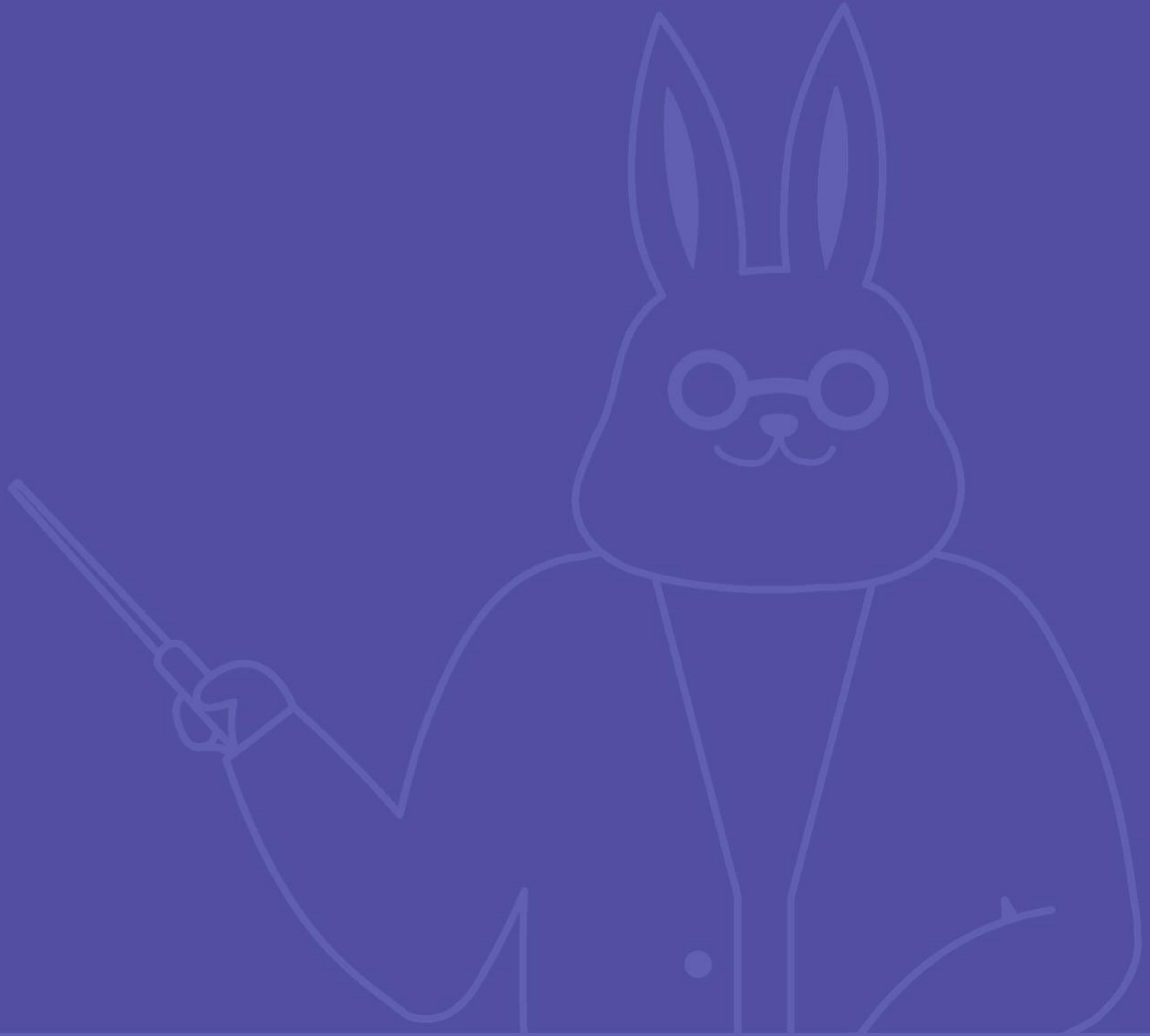
머신러닝 시작하기

03 지도학습 - 회귀



목차

01

회귀 개념 알아보기

가정해보기

아이스크림 가게를
운영하는 주인이라고 가정해보자.

판매용 아이스크림 주문 시,
예상되는 실제 판매량만큼만 주문을 원한다.

이 때 만약 **평균 기온**을 활용하여
미래 판매량을 예측할 수 있다면?

문제 정의와 해결 방안

문제 정의

- 데이터: **과거 평균 기온**(X)과 그에 따른 **아이스크림 판매량**(Y)
- 가정: 평균 기온과 판매량은 선형적인 관계를 가지고 있음
- 목표: 평균 기온에 따른 아이스크림 판매량 예측하기

해결 방안

회귀 분석 알고리즘

평균 기온(°C) (X)	아이스크림 판매량(만개) (Y)
10	40
13	52.3
20	60.5
25	80

회귀 분석이란?

데이터를 **가장 잘 설명하는 모델**을 찾아
입력값에 따른 미래 결과값을 예측하는 알고리즘

주어진 데이터

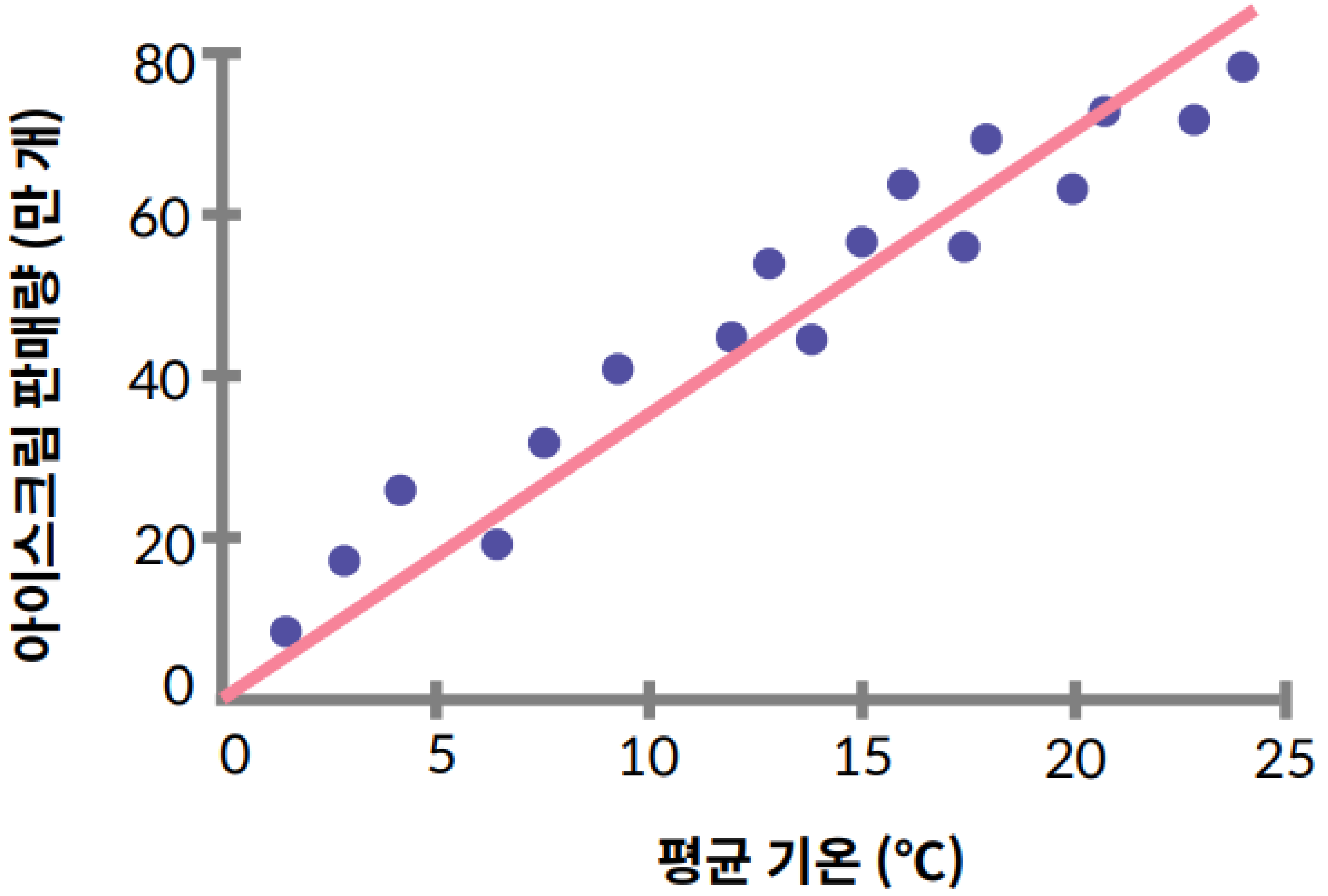
X : 평균 기온, Y : 아이스크림 판매량

가정

$Y \approx \beta_0 + \beta_1 X$ ➡ **적절한 β_0, β_1 값**을 찾자

데이터를 **가장 잘 설명하는 모델**

$Y \approx \beta_0 + \beta_1 X$



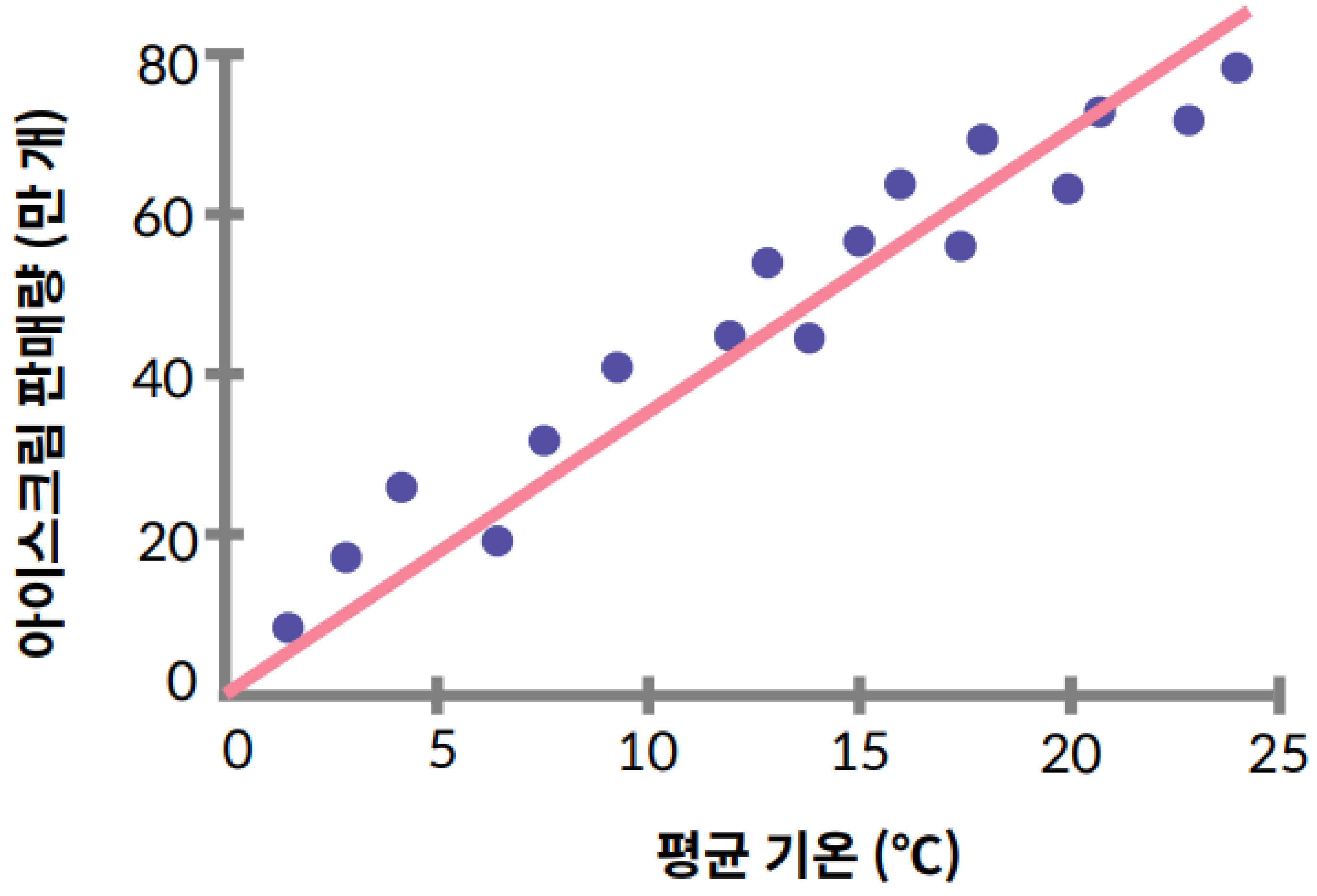
적절한 β_0, β_1 값을 찾기

완벽한 예측은 **불가능**하기에
최대한 잘 **근사**해야 한다

각 데이터의 실제 값과 모델이 예측하는 값의
차이를 최소한으로 하는 선을 찾자

단순 선형 회귀 모델을 학습하며 차이를
최소한으로 하는 선을 찾는 방법을 알아보자

$$Y \approx \beta_0 + \beta_1 X$$



02

단순 선형 회귀

단순 선형 회귀란?

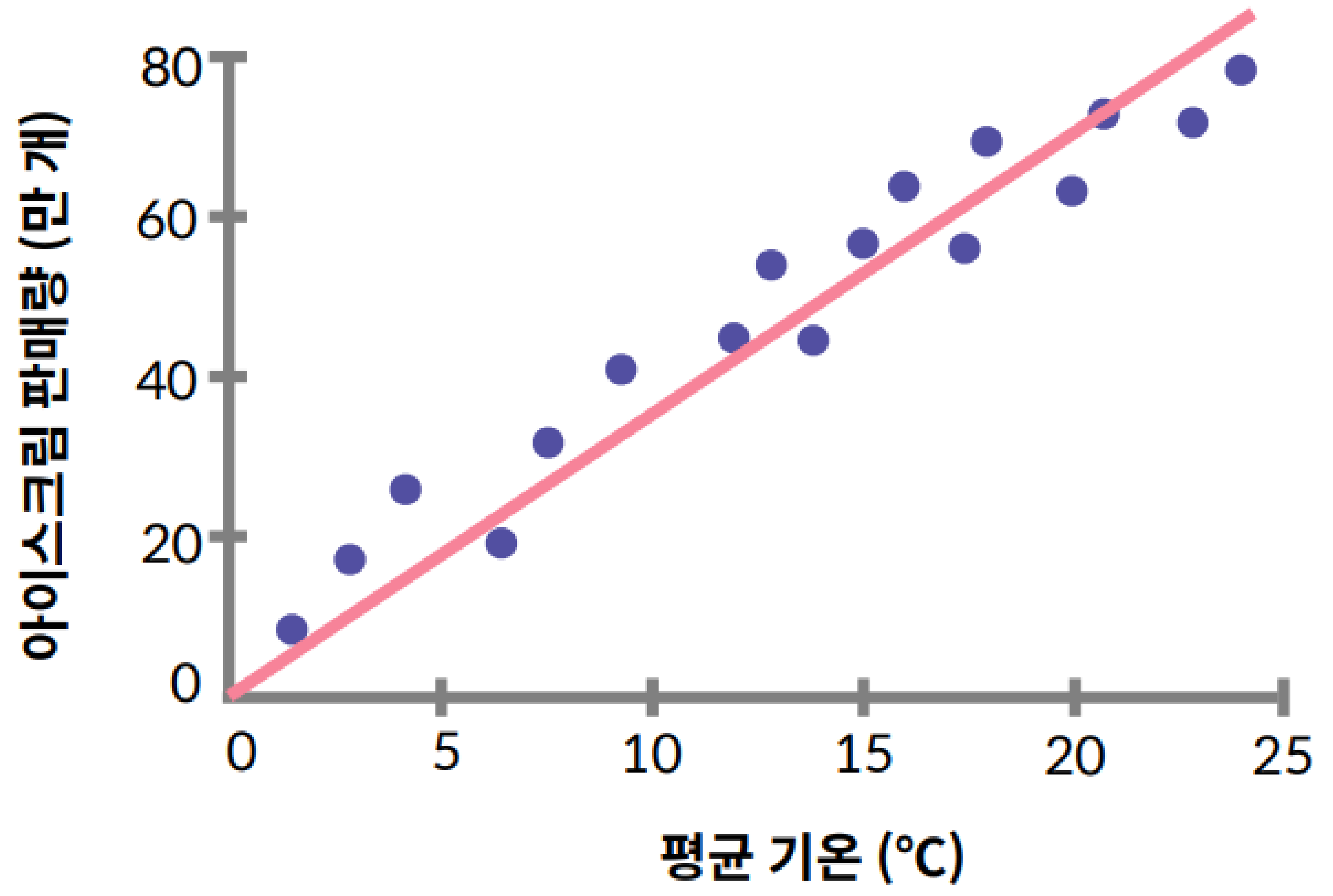
데이터를 설명하는 모델을 **직선 형태**로 가정

가정

$$Y \approx \beta_0 + \beta_1 X$$

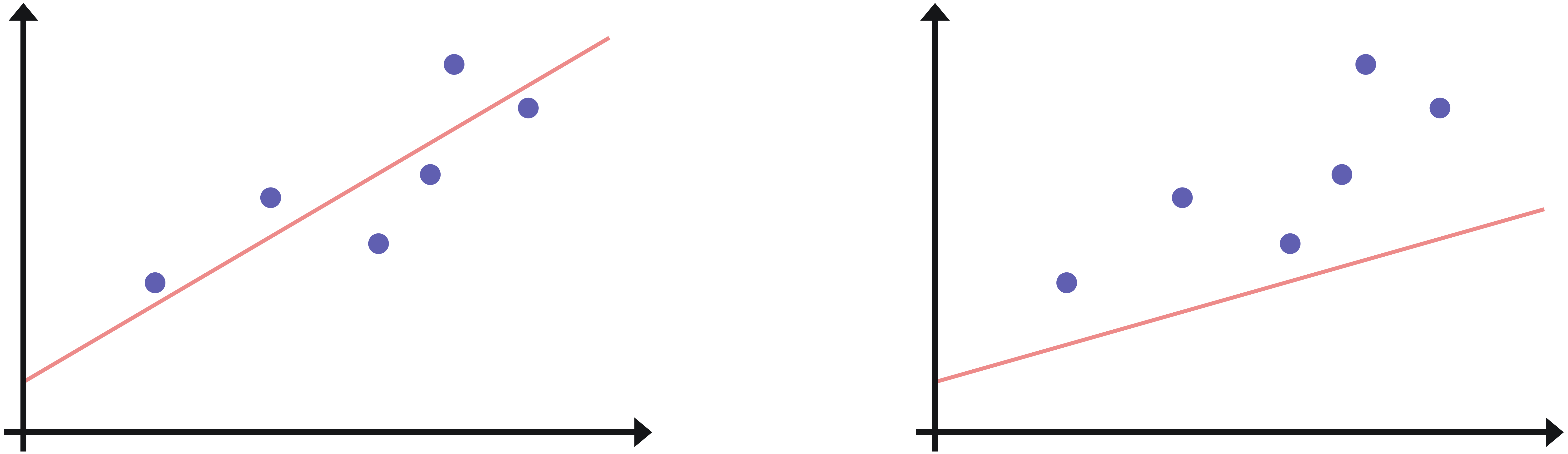
직선을 구성하는
β_0(y절편)와 **β_1(기울기)**를 구해야
함

$$Y \approx \beta_0 + \beta_1 X$$



데이터를 잘 설명한다는 것은 어떤 것일까?

실제 정답과 내가 예측한 값과의 차이가 작을수록 좋지 않을까?

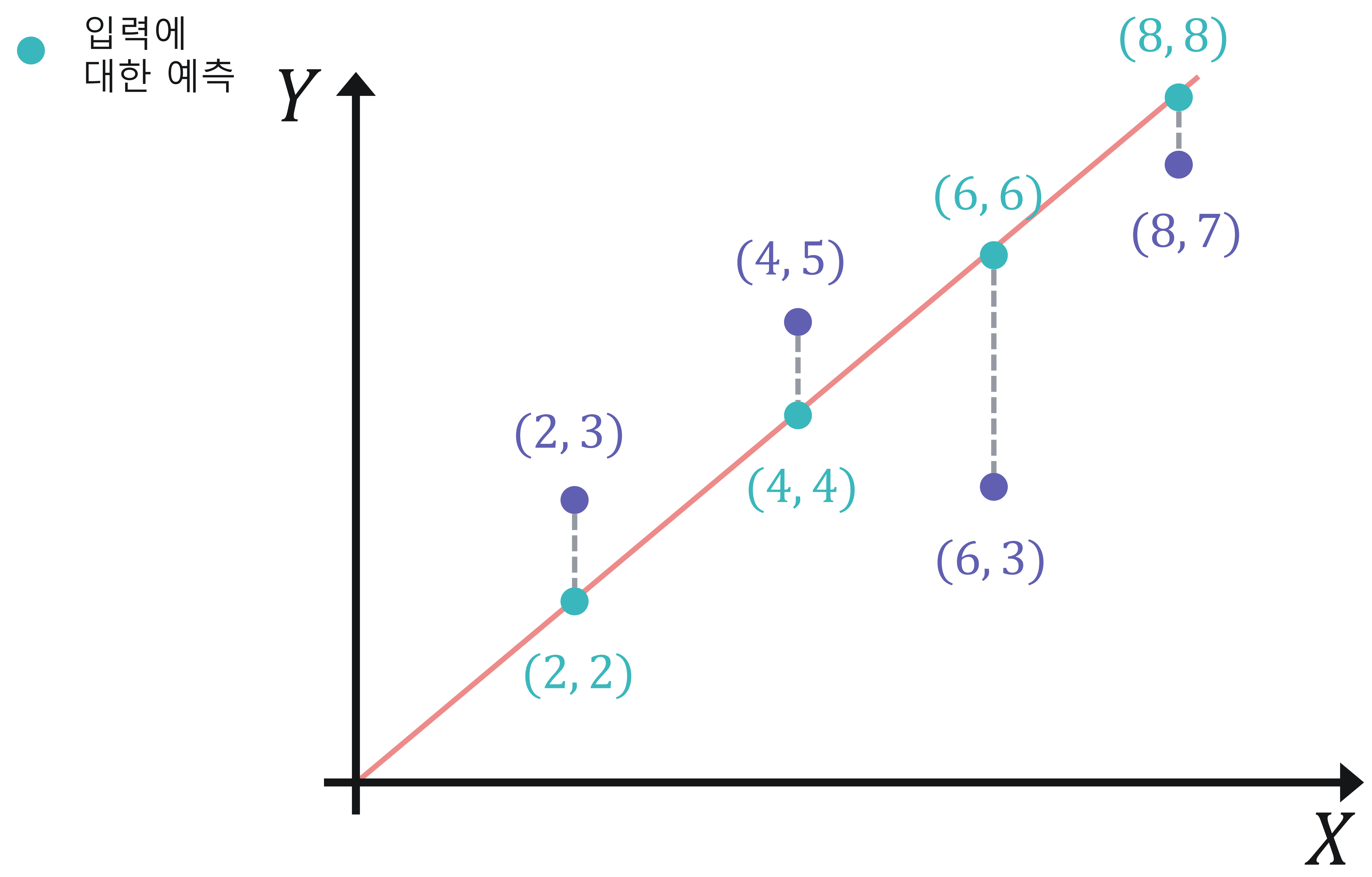


왼쪽 그래프의 차이가 오른쪽에 비해 적게 보임

데이터를 잘 설명한다는 것은 어떤 것일까?

실제 값과 예측 값의 차이를 구해보자

$$Y \approx X \quad (\beta_0 = 0,\ \beta_1 = 1)$$

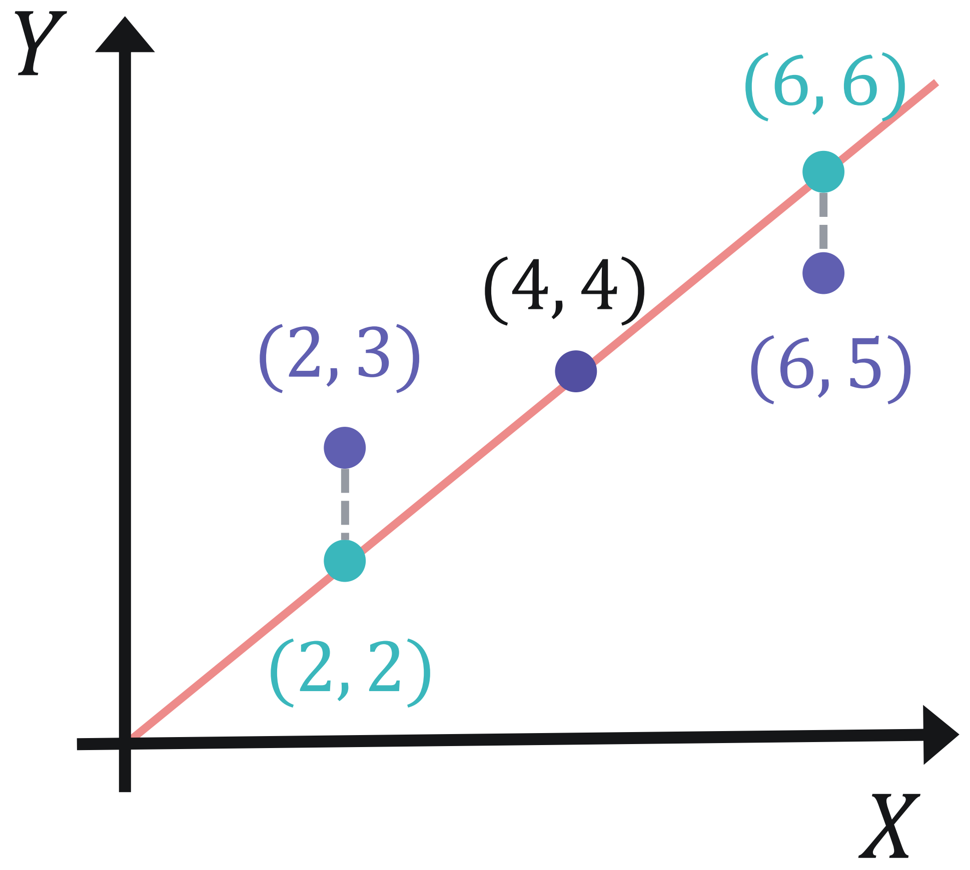


입력값	예측값	실제값	실제값 – 예측값
2	2	3	1
4	4	5	1
6	6	3	-3
8	8	7	-1
		합계	**-2**

데이터를 잘 설명한다는 것은 어떤 것일까?

실제 값과 예측 값의 차이의 합으로 비교하기에는 예외가 있다

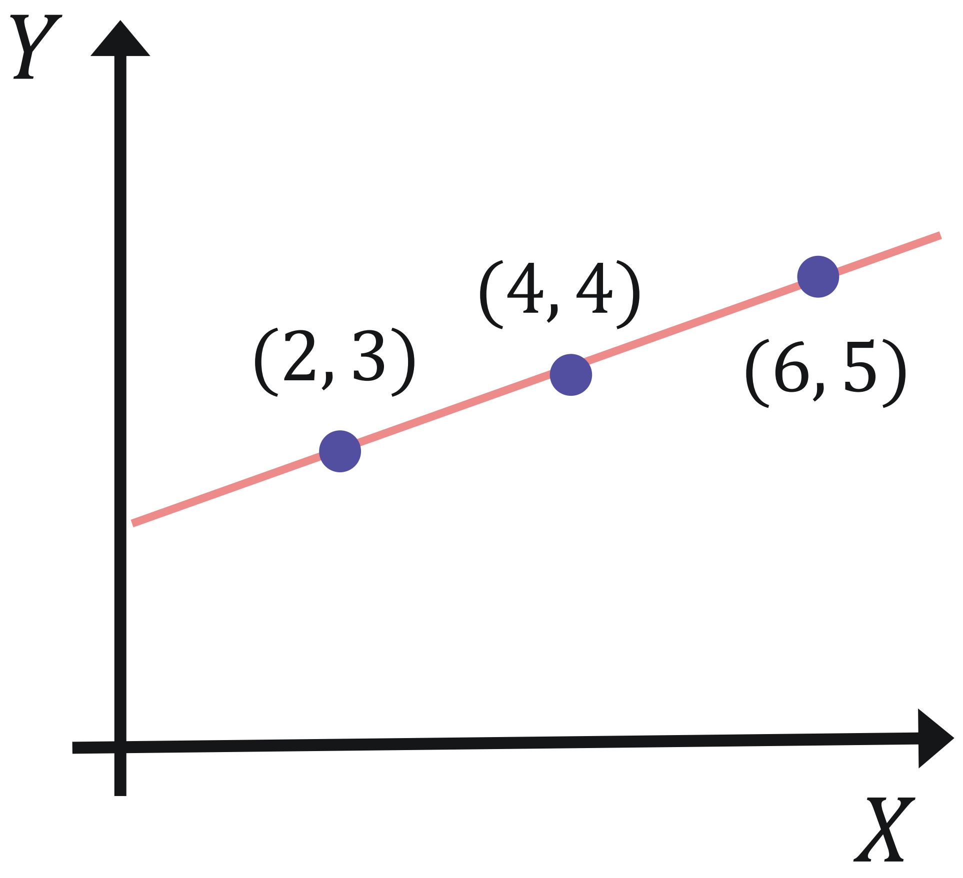
$$Y \approx X$$



입력값	예측값	실제값	실제값 – 예측값
2	2	3	1
4	4	4	0
6	6	5	-1
		합계	**0**

(4,4)에서 예측값과 실제값이 일치

$$Y \approx \frac{1}{2}X + 2$$



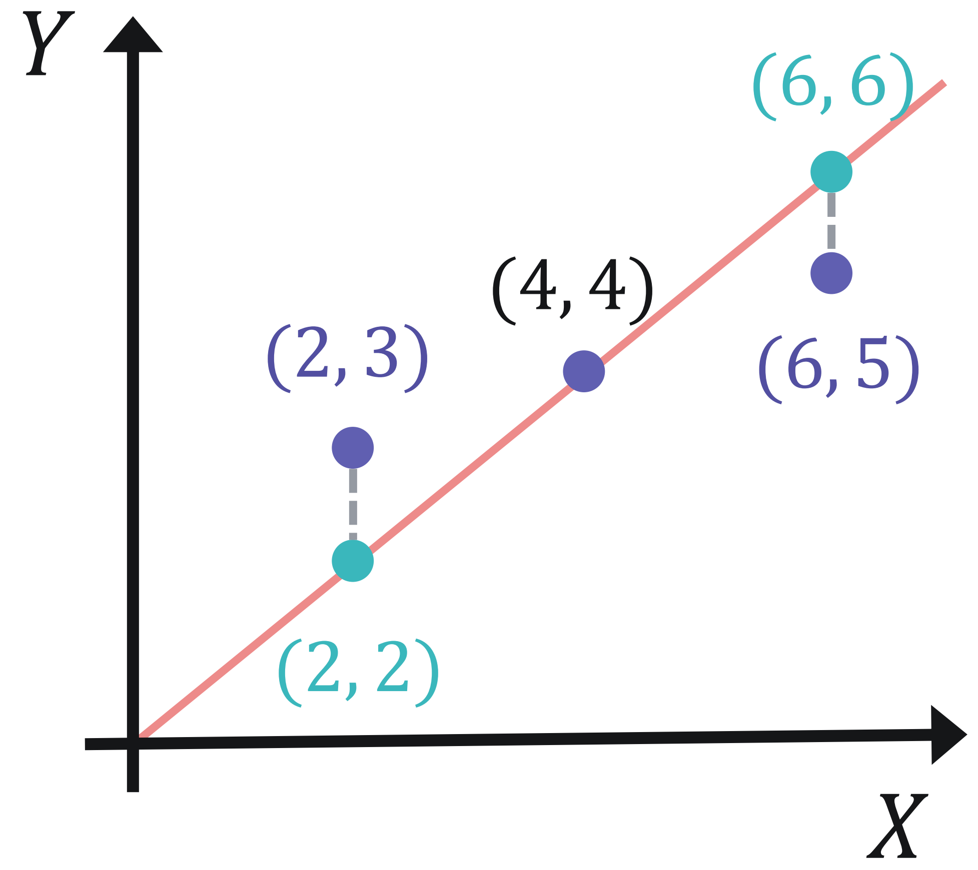
입력값	예측값	실제값	실제값 – 예측값
2	3	3	0
4	4	4	0
6	5	5	0
		합계	**0**

3가지 경우 모두에서 예측값과 실제값이 일치

데이터를 잘 설명한다는 것은 어떤 것일까?

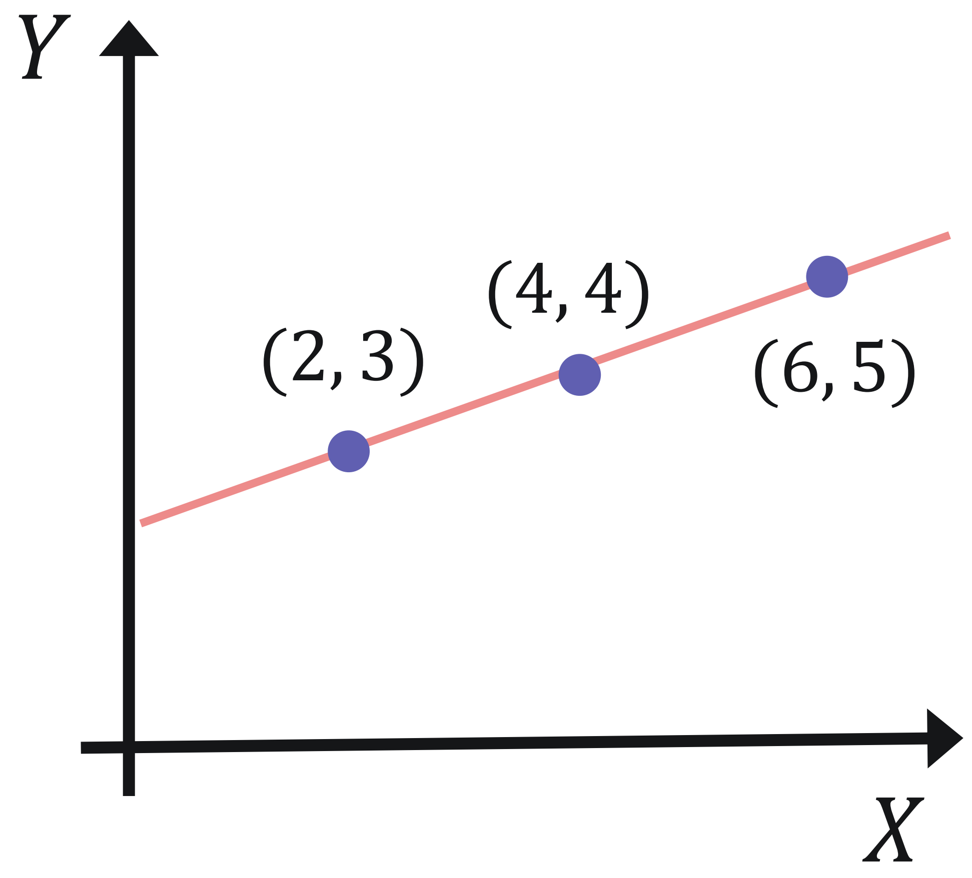
실제 값과 예측 값 **차이의 제곱의 합**으로 비교하자

$Y \approx X$



입력값	예측값	실제값	(실제값 – 예측값)²
2	2	3	1
4	4	4	0
6	6	5	1
		합계	**2**

$Y \approx \frac{1}{2}X + 2$



입력값	예측값	실제값	(실제값 – 예측값)²
2	3	3	0
4	4	4	0
6	5	5	0
		합계	**0**

Loss 함수 이해하기

실제 값과 예측 값 차이의 제곱의 합을 **Loss 함수**로 정의합니다

➡ **Loss 함수**가 작을 수록 좋은 모델이다

$$Y \approx \beta_0 + \beta_1 X$$

i 번째 데이터 $\left(x^{(i)},\ y^{(i)}\right)$ 에 대해:

입력 값: $x^{(i)}$

실제 값: $y^{(i)}$

예측 값: $\beta_0 x^{(i)} + \beta_1$

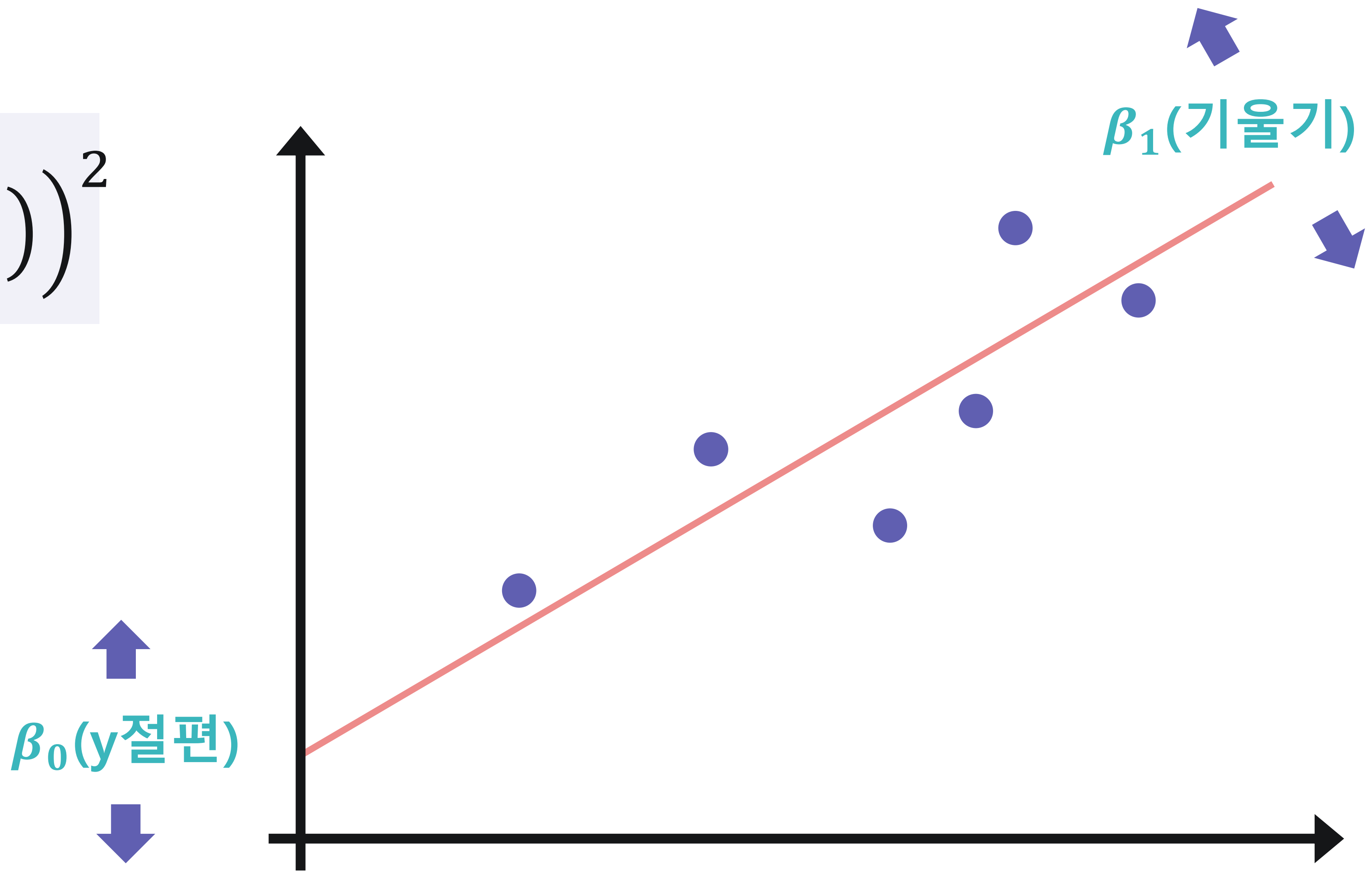
Loss 함수: $\frac{1}{N}\Sigma_i^N \left(y^{(i)} - \left(\beta_0 + \beta_1 x^{(i)}\right)\right)^2$

Loss 함수 줄이기

Loss 함수에서 주어진 값은 입력 값과 실제 값이다

➡ **β_0(y절편), β_1(기울기)** 값을 조절하여 Loss 함수의 크기를 작게 한다

Loss 함수: $\frac{1}{N}\sum_i^N\left(y^{(i)} - \left(\beta_0 + \beta_1 x^{(i)}\right)\right)^2$



Loss 함수 줄이기

Loss 함수의 크기를 작게 하는 β_0(y절편), β_1(기울기)를 찾는 방법

$$\underset{\beta_0,\beta_1}{argmin}\frac{1}{N}\sum_{i}^{N}\left(y^{(i)}-\left(\beta_0+\beta_1 x^{(i)}\right)\right)^2$$

1) **Gradient descent (경사 하강법)**
2) Normal equation (least squares)
3) Brute force search
4) …

경사 하강법

Loss 함수 값이 제일 작게 하는 절편, 기울기를 β_0^*, β_1^*라고 하자

경사 하강법은 계산 한 번으로 β_0^*, β_1^*을 구하는 것이 아니라
초기값에서 점진적으로 구하는 방식

경사 하강법

β_0, β_1 값을 Loss 함수 값이 작아지게 계속 업데이트 하는 방법

1) β_0, β_1 값을 랜덤하게 초기화

2) 현재 β_0, β_1 값으로 Loss 값 계산

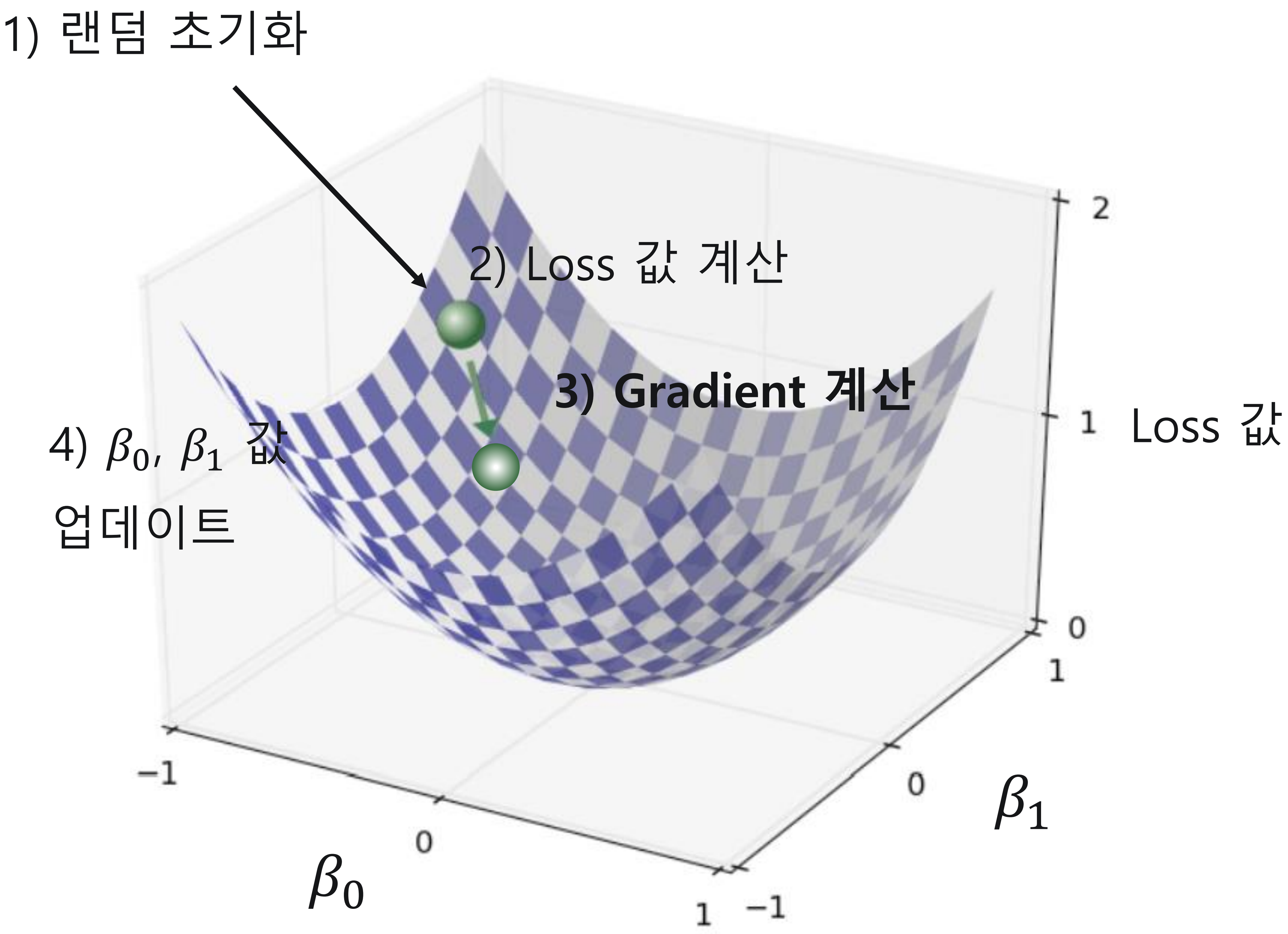
3) 현재 β_0, β_1 값을 어떻게 변화해야 Loss 값을 줄일 수 있는지 알 수 있는 **Gradient 값** 계산

4) **Gradient 값**을 활용하여 β_0, β_1 값 업데이트

5) Loss 값의 차이가 거의 없어질 때까지 **2~4번 과정을 반복**

(Loss 값과 차이가 줄어들면 Gradient 값도 작아 짐)

경사 하강법


1) 랜덤 초기화
2) Loss 값 계산
3) Gradient 계산
4) β_0, β_1 값 업데이트
Loss 값
2
1
0
β_0
-1
0
1
β_1
-1
0
1

단순 선형 회귀 과정 살펴보기

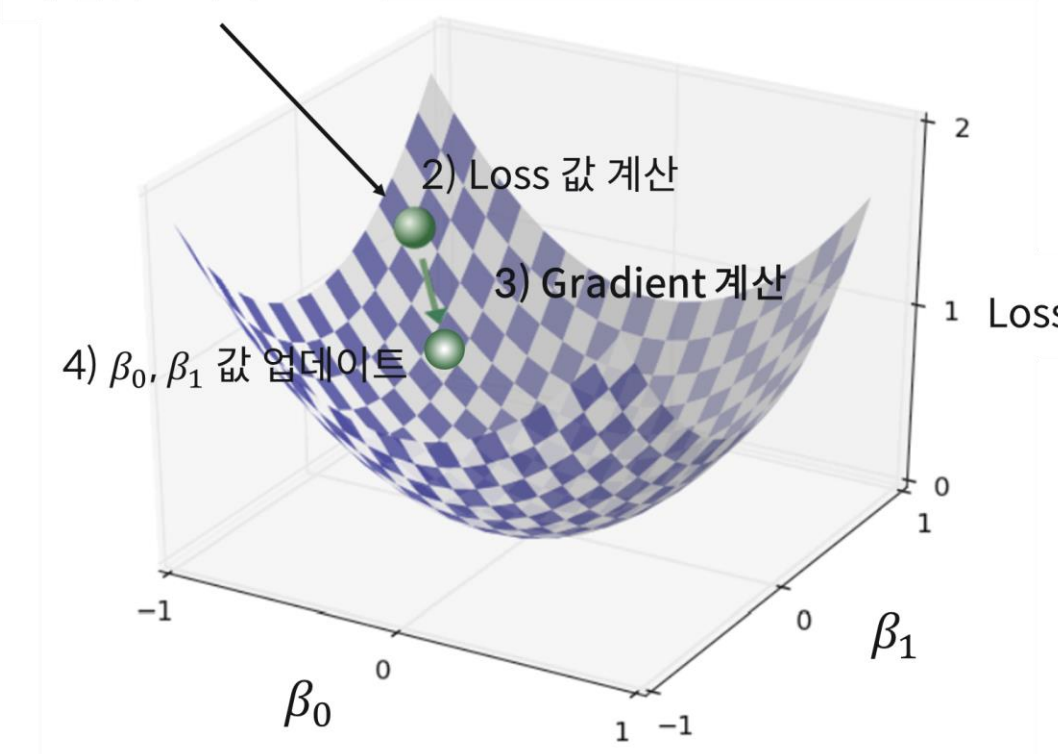
데이터 전처리

X Y

평균 기온(° C)	아이스크림 판매량(만개)
10	40
13	52.3
20	60.5
25	80

X, Y

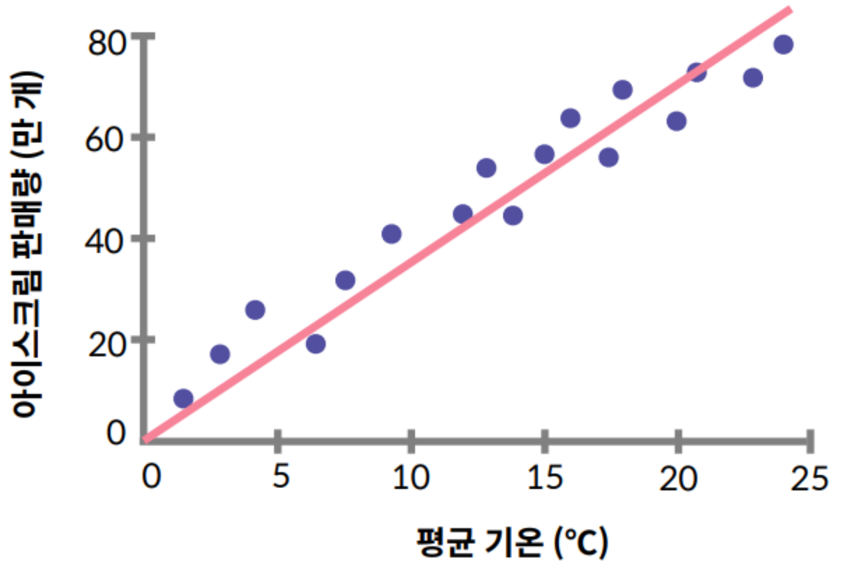
단순 선형 회귀 모델 학습 (경사 하강법)



β_0, β_1

새로운 데이터에 대한 예측

$Y \approx \beta_0 + \beta_1 X$



단순 선형 회귀 특징

- 가장 기초적이나 여전히 많이 사용되는 알고리즘
- 입력값이 1개인 경우에만 적용이 가능함
- 입력값과 결괏값의 관계를 알아보는 데 용이함
- 입력값이 결괏값에 얼마나 영향을 미치는지 알 수 있음
- 두 변수 간의 관계를 직관적으로 해석하고자 하는 경우 활용

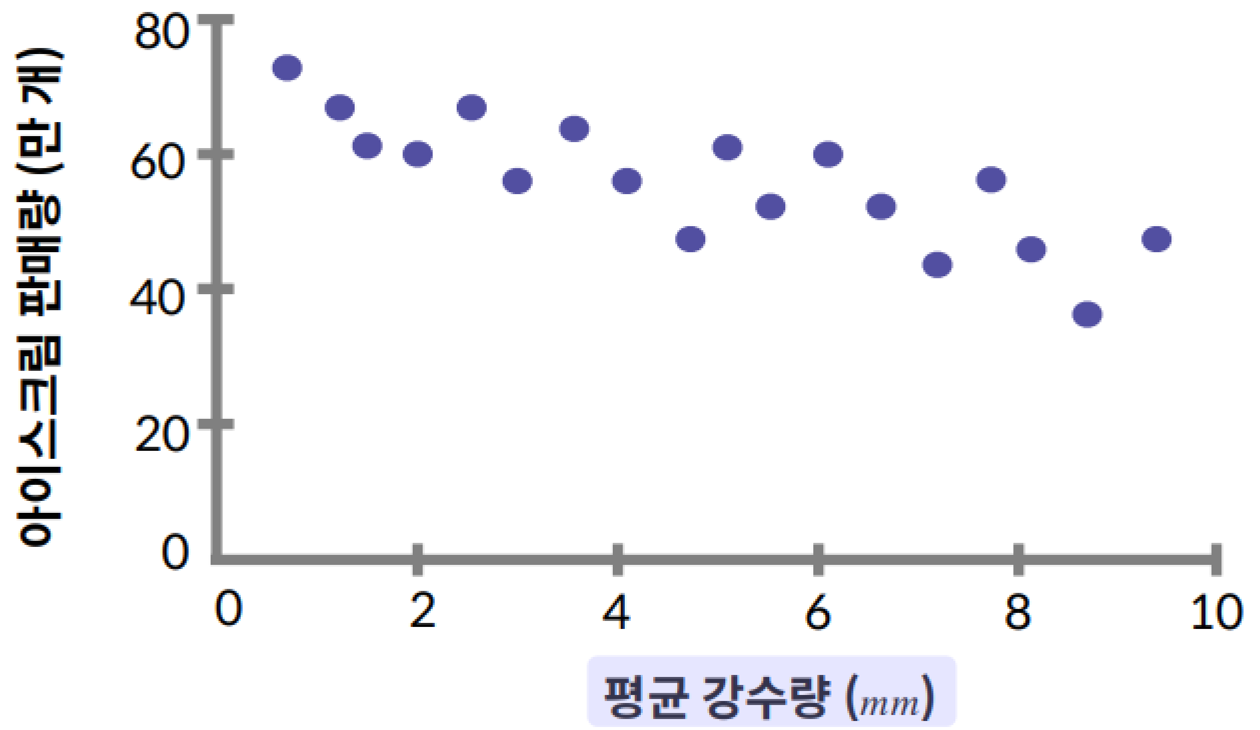
03

다중 선형 회귀

문제

만약, 입력값 X에 강수량이 추가된다면?
즉, 평균 기온과 평균 강수량에 따른 아이스크림 판매량을 예측하고자 할 때

X_1 평균 기온(° C)	X_2 평균 강수량 (mm)	Y 아이스크림 판매량(만개)
10	10	40
13	7.5	52.3
20	2	60.5
25	0	80



문제 해결하기

평균 기온과 **평균 강수량**에 따른 **아이스크림 판매량**을 예측하고자 함

X_1 X_2 Y

여러 개의 입력값(X)으로 결괏값(Y)을 예측하고자 하는 경우 → **다중 선형 회귀 (Multiple Linear Regression)**

다중 선형 회귀 모델 이해하기

입력값 X가 여러 개(2개 이상)인 경우 활용할 수 있는 회귀 알고리즘

각 **개별 X_i**에 해당하는 **최적의 β_i**를 찾아야 함

다중 선형 회귀 모델

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + +\beta_M X_M$$

평균 기온과 **평균 강수량**에 따른 **아이스크림 판매량**을 예측하고자 함

X_1 X_2 Y

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

다중 선형 회귀 모델의 Loss 함수

단순 선형 회귀와 마찬가지로 **Loss 함수**는 **입력값과 실제값 차이의 제곱의 합**으로 정의합니다

➡ **마찬가지로 $\boldsymbol{\beta_0, \beta_1, \beta_2, \ldots, \beta_M}$**값을 조절하여 Loss 함수의 크기를 작게 합니다

Loss 함수: $\frac{1}{N}\sum_i^N \left(y^{(i)} - \left(\beta_0 + \beta_1 {x_1}^{(i)} + \beta_2 {x_2}^{(i)} + \cdots + \beta_M {x_M}^{(i)}\right)\right)^2$

다중 선형 회귀 모델의 Loss 함수

평균 기온과 평균 강수량으로 아이스크림 판매량 예측 예시

Loss 함수: $\frac{1}{N}\sum_{i}^{N}\left(y^{(i)}-\left(10+2{x_1}^{(i)}+2{x_2}^{(i)}\right)\right)^2$

입력값 1 (평균 기온)	입력값 2 (평균 강수량)	예측값	실제값 (아이스크림 판매량)	(실제값 – 예측값)²
10	10	50	40	100
13	7.5	51	52.3	1.69
20	2	54	60.5	42.25
25	0	60	80	400
			합계	**543.94**

다중 선형 회귀 모델의 경사 하강법

$\beta_0, \beta_1, \beta_2, \ldots, \beta_M$ 값을 Loss 함수 값이 작아지게 계속 업데이트 하는 방법

1) $\beta_0, \beta_1, \beta_2, \ldots, \beta_M$ 값을 랜덤하게 초기화

2) 현재 $\beta_0, \beta_1, \beta_2, \ldots, \beta_M$값으로 Loss 값 계산

3) 현재 $\beta_0, \beta_1, \beta_2, \ldots, \beta_M$ 값을 어떻게 변화해야 Loss 값을 줄일 수 있는지 알 수 있는 **Gradient 값** 계산

4) **Gradient 값**을 활용하여 $\beta_0, \beta_1, \beta_2, \ldots, \beta_M$ 값 업데이트

5) Loss 값의 차이가 거의 없어질 때까지 2~4번 과정을 반복
(Loss 값과 차이가 줄어들면 Gradient 값도 작아 짐)

다중 선형 회귀 모델의 경사 하강법 결과 예시

평균 기온과 평균 강수량으로 아이스크림 판매량 예측 예시

$$Y \approx -49.48 + 5.17X_1 + 4.05X_2$$

입력값 1 (평균 기온)	입력값 2 (평균 강수량)	예측값	실제값 (아이스크림 판매량)	(실제값 – 예측값)^2
10	10	42.79	40	7.78
13	7.5	48.16	52.3	17.14
20	2	62.05	60.5	4
25	0	79.79	80	0.04
			합계	**42.89**

다중 선형 회귀 특징

- 여러 개의 입력값과 결괏값 간의 관계 확인 가능
- 어떤 입력값이 결괏값에 어떠한 영향을 미치는지 알 수 있음
- 여러 개의 입력값 사이 간의 **상관 관계***가 높을 경우 결과에 대한 신뢰성을 잃을 가능성이 있음

상관 관계

- 두 가지 것의 한쪽이 변화하면 다른 한쪽도 따라서 변화하는 관계

04

회귀 평가 지표

회귀 알고리즘 평가

어떤 모델이 좋은 모델인지를 어떻게 평가할 수 있을까?
목표를 얼마나 잘 달성했는지 정도를 평가해야 함

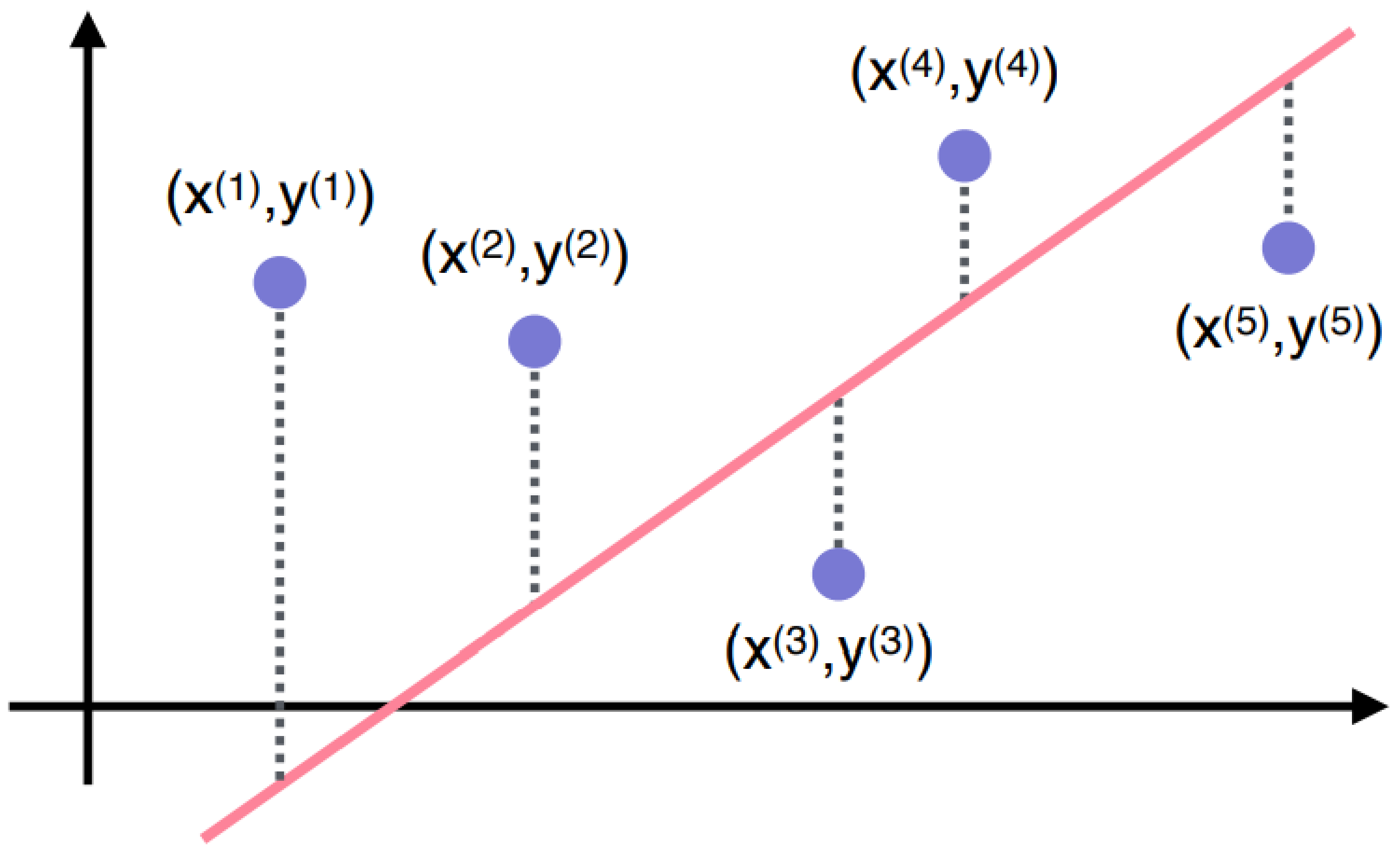
실제 값과 모델이 예측하는 값의 **차이**에 기반한 평가 방법 사용

예시

$$RSS, MSE, MAE, MAPE, R^2$$

RSS – 단순 오차

1. 실제 값과 예측 값의 단순 오차 제곱 합
2. 값이 작을수록 모델의 성능이 높음
3. 전체 데이터에 대한 실제 값과 예측하는 값의 오차 제곱의 총합



RSS

$$RSS = \sum_{i}^{N}\left(y^{(i)} - \left(\beta_0 + \beta_1 x^{(i)}\right)\right)^2$$

RSS 특징

- 가장 간단한 평가 방법으로 직관적인 해석이 가능함
- 그러나 오차를 그대로 이용하기 때문에 입력 값의 **크기에 의존적**임
- 절대적인 값과 비교가 불가능함

MSE, MAE – 절대적인 크기에 의존한 지표

- MSE(Mean Squared Error)

평균 제곱 오차, RSS 에서 데이터 수 만큼 나눈 값

작을수록 모델의 성능이 높다고 평가할 수 있음.

$$MSE = \frac{1}{N}\sum_{i}^{N}\left(y^{(i)} - \left(\beta_0 + \beta_1 x^{(i)}\right)\right)^2$$

MSE, MAE – 절대적인 크기에 의존한 지표

- **MAE(Mean Absolute Error)**

평균 절댓값 오차, 실제 값과 예측 값의 오차의 절대값의 평균
작을수록 모델의 성능이 높다고 평가할 수 있음.

$$MAE = \frac{1}{N}\sum_{i}^{N}\left|y^{(i)} - \left(\beta_0 + \beta_1 x^{(i)}\right)\right|$$

MSE, MAE 특징

- **MSE:** 이상치(Outlier) 즉, 데이터들 중 크게 떨어진 값에 민감함
- **MAE:** 변동성이 큰 지표와 낮은 지표를 같이 예측할 시 유용
- 가장 간단한 평가 방법들로 직관적인 해석이 가능함
- 그러나 평균을 그대로 이용하기 때문에 입력 값의 크기에 의존적임
- 절대적인 값과 비교가 불가능함

R^2 (결정 계수)

회귀 모델의 설명력을 표현하는 지표

1에 가까울수록 높은 성능의 모델이라고 해석할 수 있음

$$R^2 = 1 - \frac{RSS}{TSS}$$

TSS는 데이터 평균 값($\bar{y}$)과 실제 값($y^{(i)}$) 차이의 제곱

$$TSS = \sum_{i}^{N} \left(y^{(i)} - \bar{y}\right)^2 \qquad \bar{y} = \frac{1}{N}\sum_{i}^{N} y^{(i)}$$

R^2 특징

- 오차가 없을수록 1에 가까운 값을 가짐
- 값이 0인 경우, 데이터의 평균 값을 출력하는 직선 모델을 의미함
- 음수 값이 나온 경우, 평균값 예측 보다 성능이 좋지 않음