

인공지능을 위한 머신러닝 알고리즘

2. 선형 회귀 모델

CONTENTS

1

선형 회귀 모델

2

파라미터 예측: 최소 제곱 방법

3

선형 회귀 모델로는 안 풀리는 문제들

학습 목표

- 선형 회귀의 분류 원리에 대해 이해할 수 있다.
- 선형 회귀의 모델 파라미터를 계산할 수 있다.
- 모델이 가정하고 있는 선형성에 대해서 이해할 수 있다.

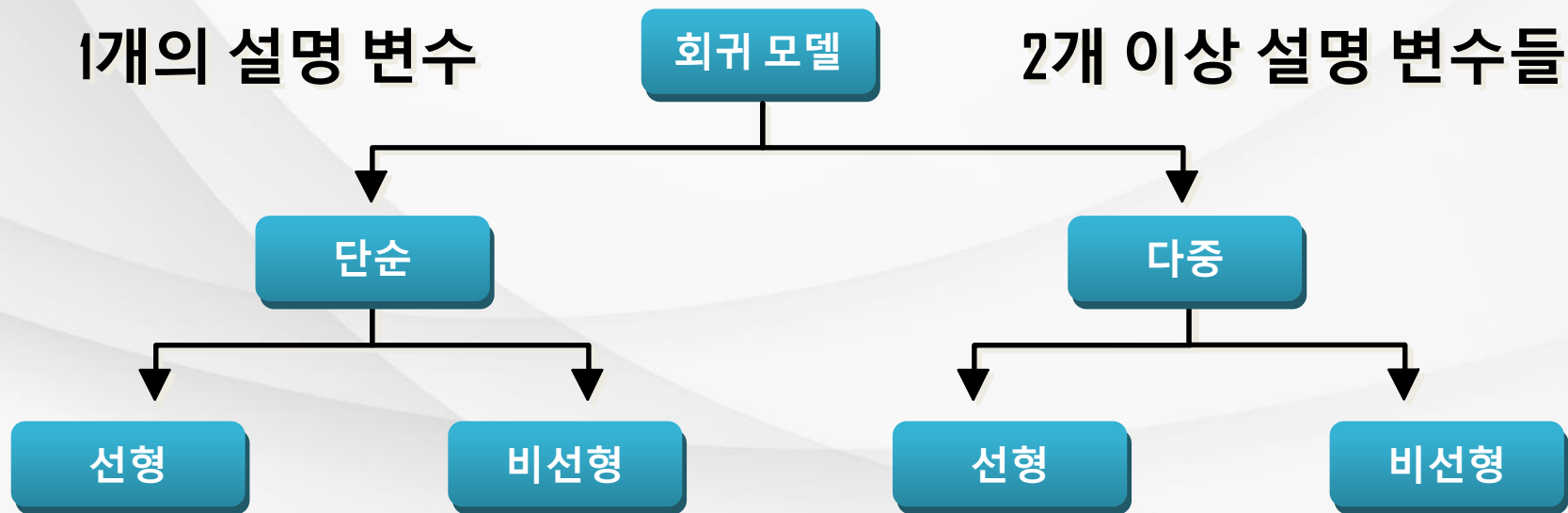


1. 선형 회귀 모델

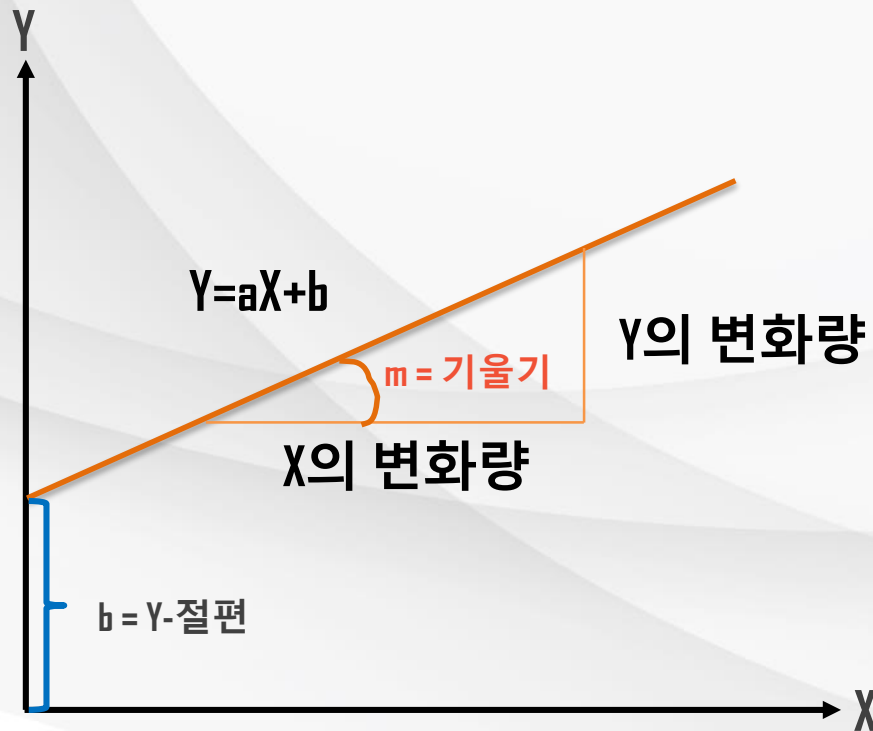
■ 회귀 모델의 정의

- ◉ 한 개의 종속 변수(**dependent variable**)와 설명 변수들(**explanatory variable(s)**)과의 관계를 모델링
- ◉ 관계를 정의하기 위해 방정식 사용
 - 수치적(**numerical**) 종속 변수
 - 한 개 또는 그 이상의 수치적 설명 변수
- ◉ 예측 & 추정 시에 사용

회귀 모델의 종류들



■ 선형성이란



예> $Y=1/2X+3$

■ 변수들 사이 관계 - 선형 함수

The diagram shows the linear regression equation $Y_i = \beta_0 + \beta_i X_i + \varepsilon_i$ enclosed in a rounded rectangle. Arrows point from descriptive labels to each term in the equation: Y_i is labeled '종속 (응답) 변수' (Dependent variable) with the example '예> 와인 등급' (e.g., wine grade); β_0 is labeled 'Y절편' (Y-intercept); β_i is labeled '기울기' (Slope); X_i is labeled '독립 (설명) 변수' (Independent variable) with the example '예> 날씨, 토양, 포도의 품질' (e.g., weather, soil, grape quality); and ε_i is labeled '무작위 에러 (노이즈)' (Random error/noise).

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

종속 (응답) 변수
예> 와인 등급

독립 (설명) 변수
예> 날씨, 토양, 포도의 품질

Y절편

기울기

무작위 에러 (노이즈)

회귀 모델의 예측 대상

모집단 (참값)



회귀 모델의 예측 대상

모집단 (참값)

무작위 샘플 (관측 값)

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

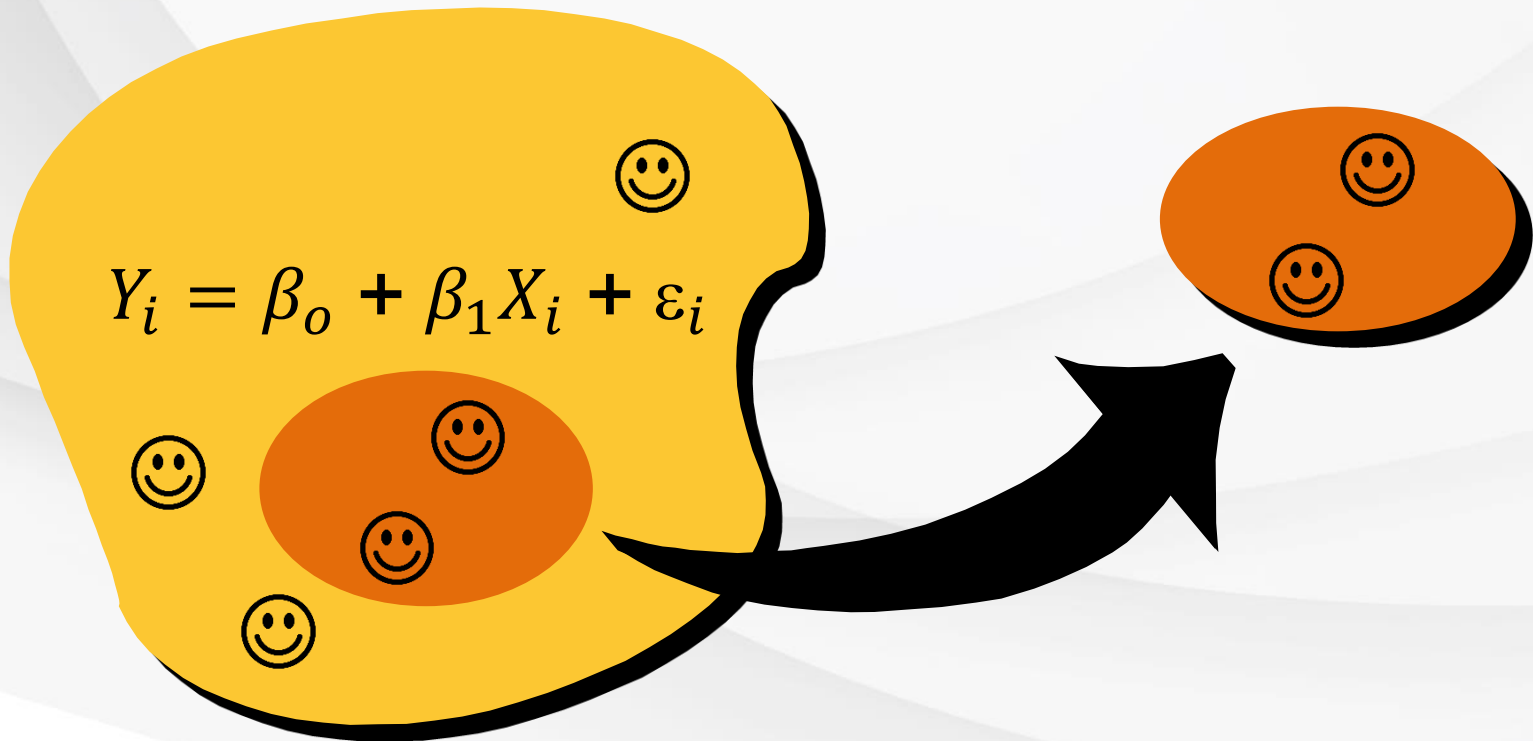
실제 데이터 생성 규칙

회귀 모델의 예측 대상

모집단 (참값)

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

무작위 샘플 (관측 값)



회귀 모델의 예측 대상

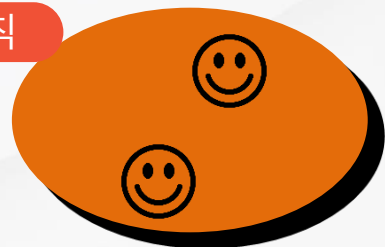
모집단 (참값)

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

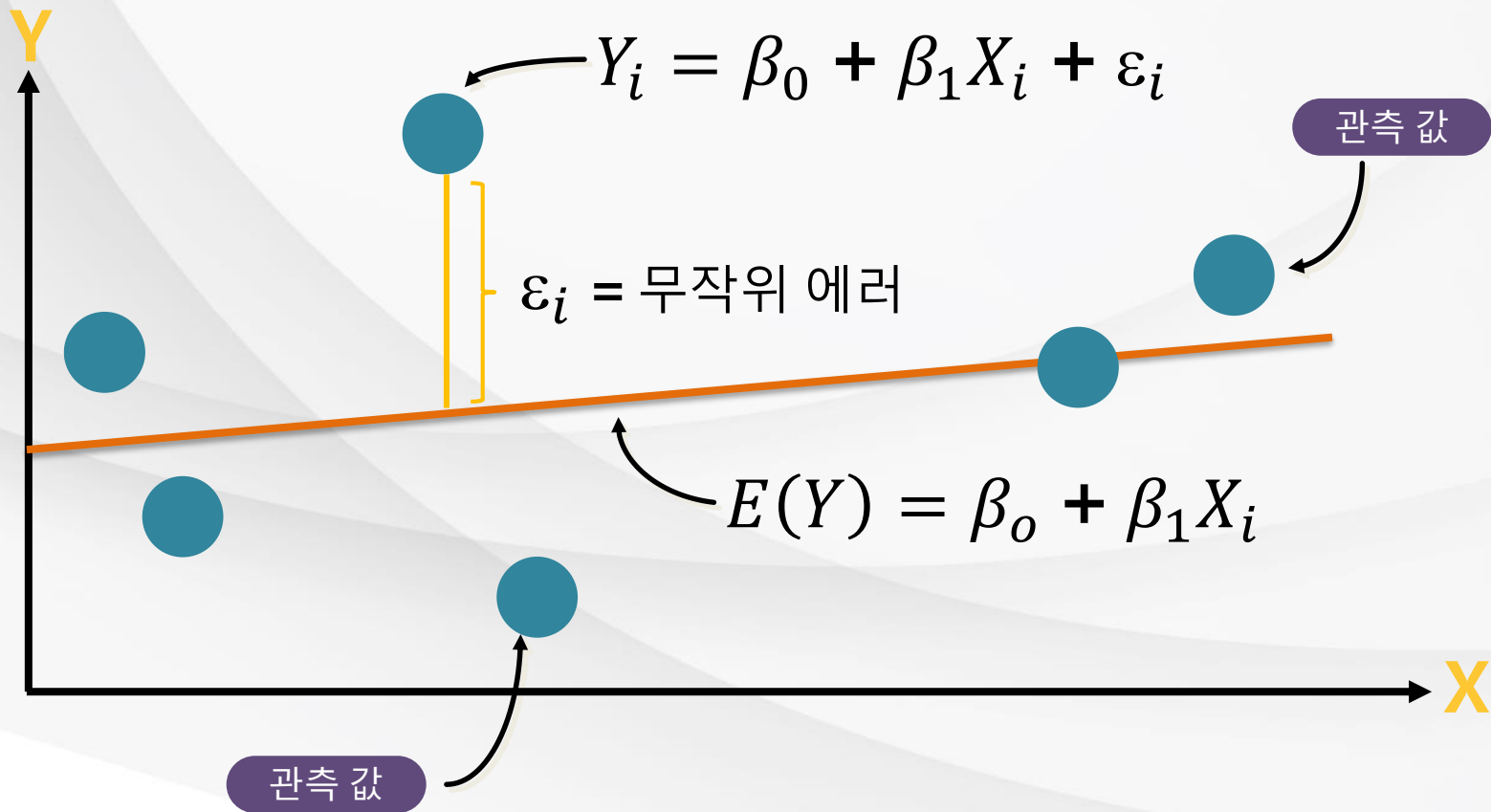

무작위 샘플 (관측 값)

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\varepsilon}_i$$

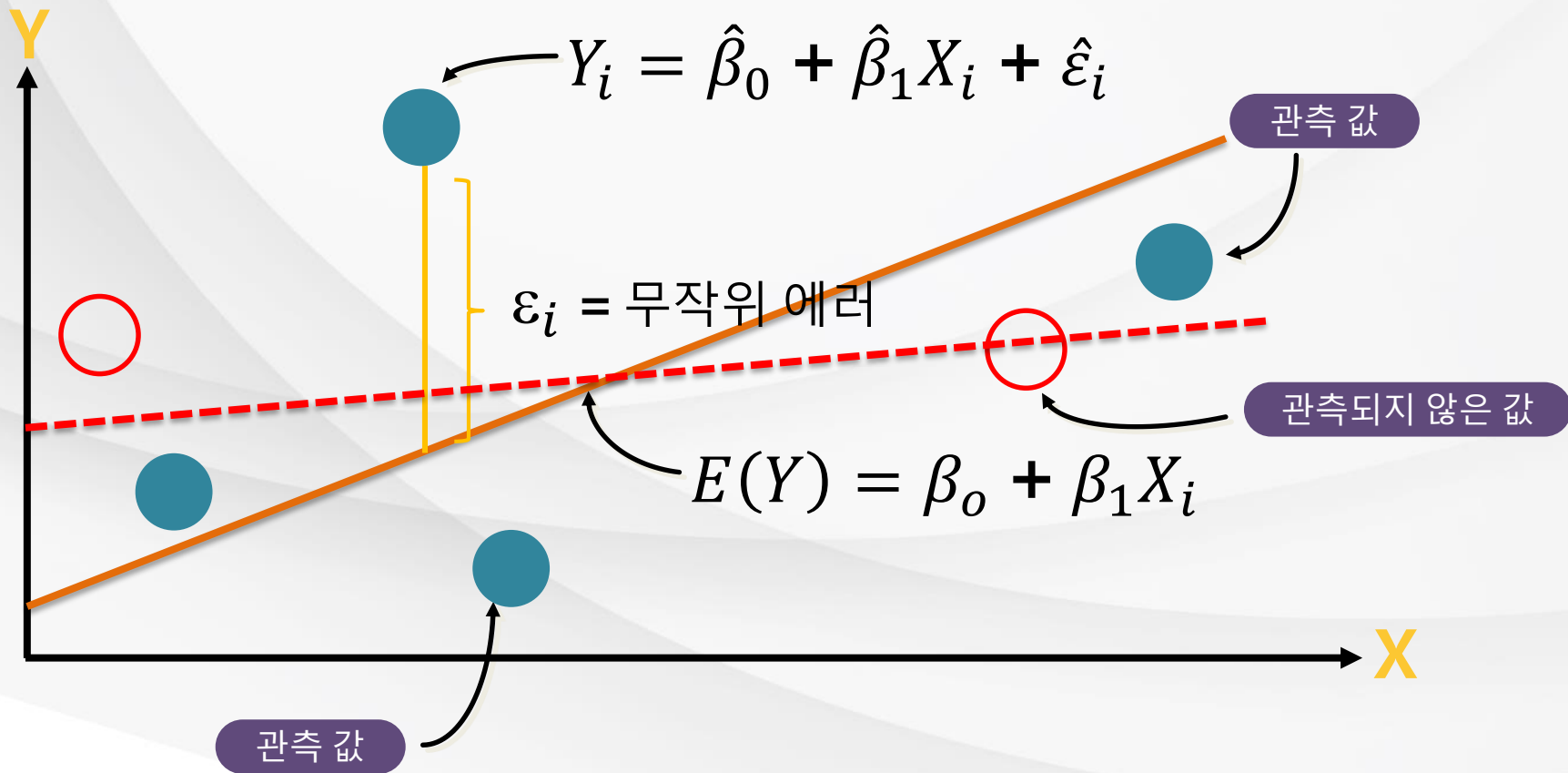
예측한 생성 규칙



■ 선형 회귀 모델의 확률적 관점



■ 선형 회귀 모델의 확률적 관점





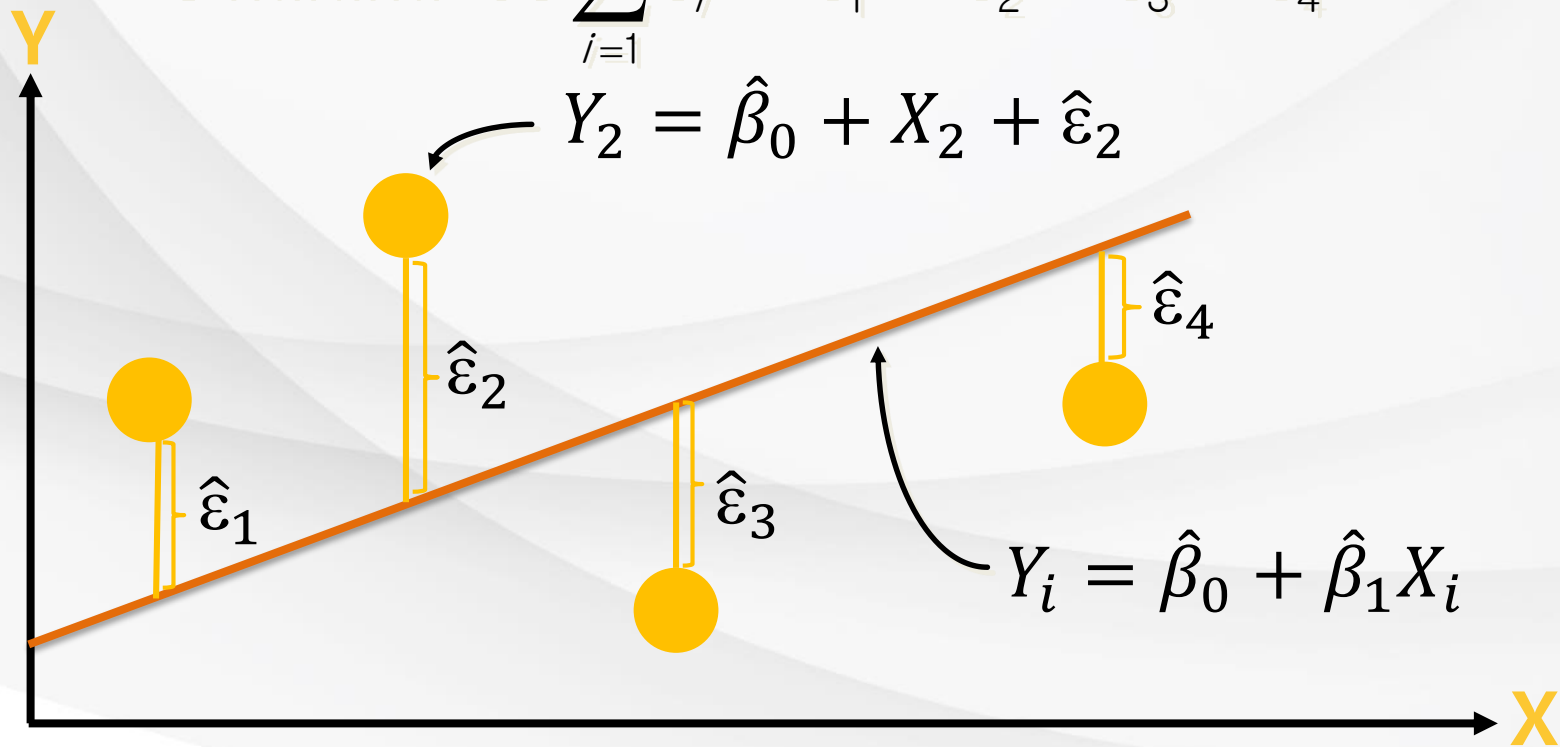
2. 파라미터 예측: 최소 제공 방법

■ 최소 제곱

- 최적의 모델은 실제 Y 값과 예측된 Y 값의 차이 (에러)가 최소가 되어야 함
- 에러의 값은 무조건 양수이어야 하므로 제곱을 시킴
- 최소 제곱 방법은 에러의 제곱의 합 (**Sum of the Squared Errors, SSE**)을 최소화 시킴

■ 최소 제곱

LS minimizes $\sum_{i=1}^n \hat{\varepsilon}_i^2 = \hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2 + \hat{\varepsilon}_3^2 + \hat{\varepsilon}_4^2$



■ 최소 제곱의 해

❖ 예측 방정식

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

❖ 기울기

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

❖ Y-절편

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

■ Y 절편 구하기

$$\sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\begin{aligned} 0 &= \frac{\partial \sum \varepsilon_i^2}{\partial \beta_0} = \frac{\partial \sum (y_i - \beta_0 - \beta_1 x_i)^2}{\partial \beta_0} \\ &= -2(n\bar{y} - n\beta_0 - n\beta_1 \bar{x}) \end{aligned}$$

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

기울기 구하기

$$0 = \frac{\partial \sum \varepsilon_i^2}{\partial \beta_1} = \frac{\partial \sum (y_i - \beta_0 - \beta_1 x_i)^2}{\partial \beta_1}$$

$$= -2 \sum x_i (y_i - \beta_0 - \beta_1 x_i)$$

$$= -2 \sum x_i (y_i - \bar{y} + \beta_1 \bar{x} - \beta_1 x_i)$$

$$\beta_1 \sum x_i (x_i - \bar{x}) = \sum x_i (y_i - \bar{y})$$

$$\beta_1 \sum (x_i - \bar{x})(x_i - \bar{x}) = \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$


■ 기울기와 y-절편의 의미

1. 기울기 ($\hat{\beta}_1$) - 추정된 Y 는 X 가 1 단위 증가할 때마다 $\hat{\beta}_1$ 만큼 변화

- 만약 $\hat{\beta}_1 = 2$ 인 경우, Y 는 X 가 1 단위 증가할 때마다 2씩 증가

2. Y-절편 ($\hat{\beta}_0$) - $X = 0$ 인 경우 Y 의 평균 값

- 만약 $\hat{\beta}_0 = 4$ 인 경우, X 가 0일 때 Y 의 평균값은 4

A person's hands are shown holding a smartphone, with the screen glowing. The background is dark with out-of-focus, colorful bokeh lights in shades of yellow, orange, and blue. A semi-transparent dark blue banner is at the bottom, containing a yellow stylized 'L' icon and the title text.

3. 선형 회귀 모델로는 안 풀리는 문제 들

3. 선형 회귀 모델로는 안 풀리는 문제

들

■ 무엇이 더 좋은 모델일까?

❖ 예> 프로그래밍 숙제의 성공 여부 예측

성공률



프로그래밍 경력 (months)

성공률



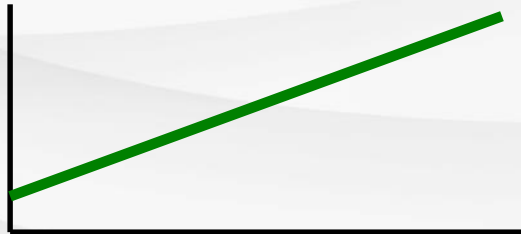
프로그래밍 경력 (months)

성공률



프로그래밍 경력 (months)

성공률



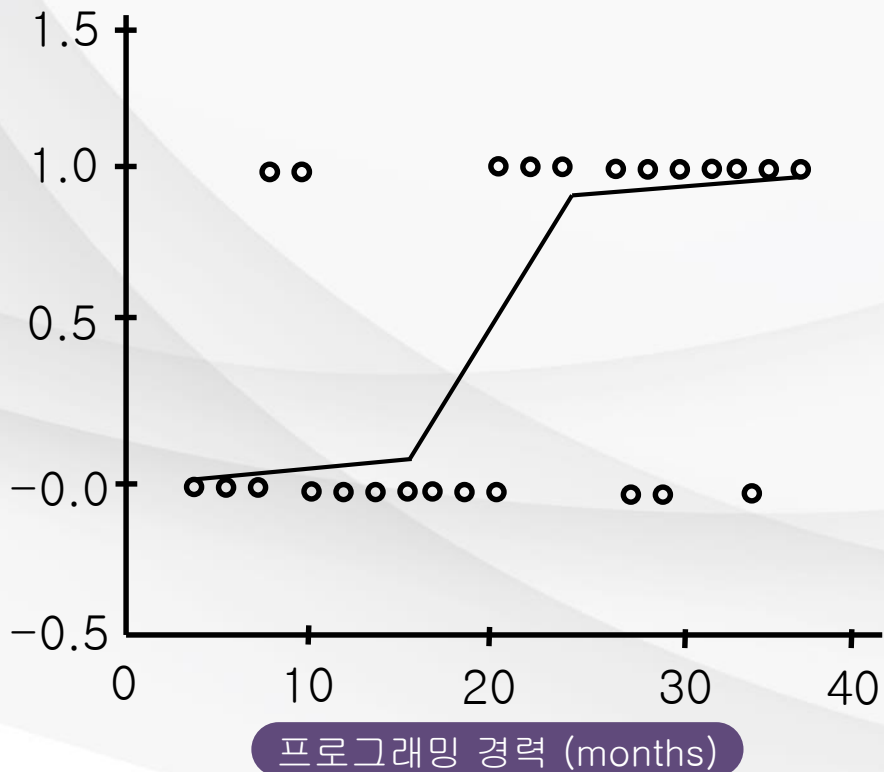
프로그래밍 경력 (months)

3. 선형 회귀 모델로는 안 풀리는 문제

들

■ 선형성의 한계

❖ 예> 프로그래밍 숙제의 성공 여부 예측





학습정리

지금까지 [선형 회귀 모델]에 대해서 살펴보았습니다.

선형 회귀 모델

한 개의 종속 변수와 다수의 설명 변수들 사이의 관계를 선형 방정식으로 모델링

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

파라미터 예측: 최소 제곱 방법

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

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

선형 문제로는 풀리지 않는 문제들

종속변수와 설명변수 사이 비선형 관계 존재