데이터 사이의 관계 찾기

2025년 5월 10일

서울대학교 통계학과
임채영

시작하기

데이터(자료)란?

- 키, 몸무게 (연속 데이터)
- 교통사고 횟수, 암환자 수 (가산 데이터)
- 성별, 자동차 색상 (범주형 데이터)

최근에는

- 음성, 사진, 동영상
- 네트워크
- 소설, 역사책

데이터를 다루는 학문 – 통계학

- 알고 싶은 질문에 답하기 위해 정보가 필요함- 데이터 추출(획득)
- 추출한 데이터의 특징을 이해 – 데이터 분석
- 분석결과를 바탕으로 질문에 대한 답/결론을 이끌어냄 – 데이터기반 의사결정

통계학자? 데이터과학자?

- 데이터과학(Data Science)이란?
- 데이터 저장/관리- 컴퓨터과학/전산학
- 데이터 추출방법/분석방법 – 통계학
- 분석 알고리즘 구현 – 컴퓨터과학/전산학
- 인사이트 도출 – 각 분야 지식 (Domain knowledge)

데이터 사이의 관계 찾기

데이터사이의 관계 찾기

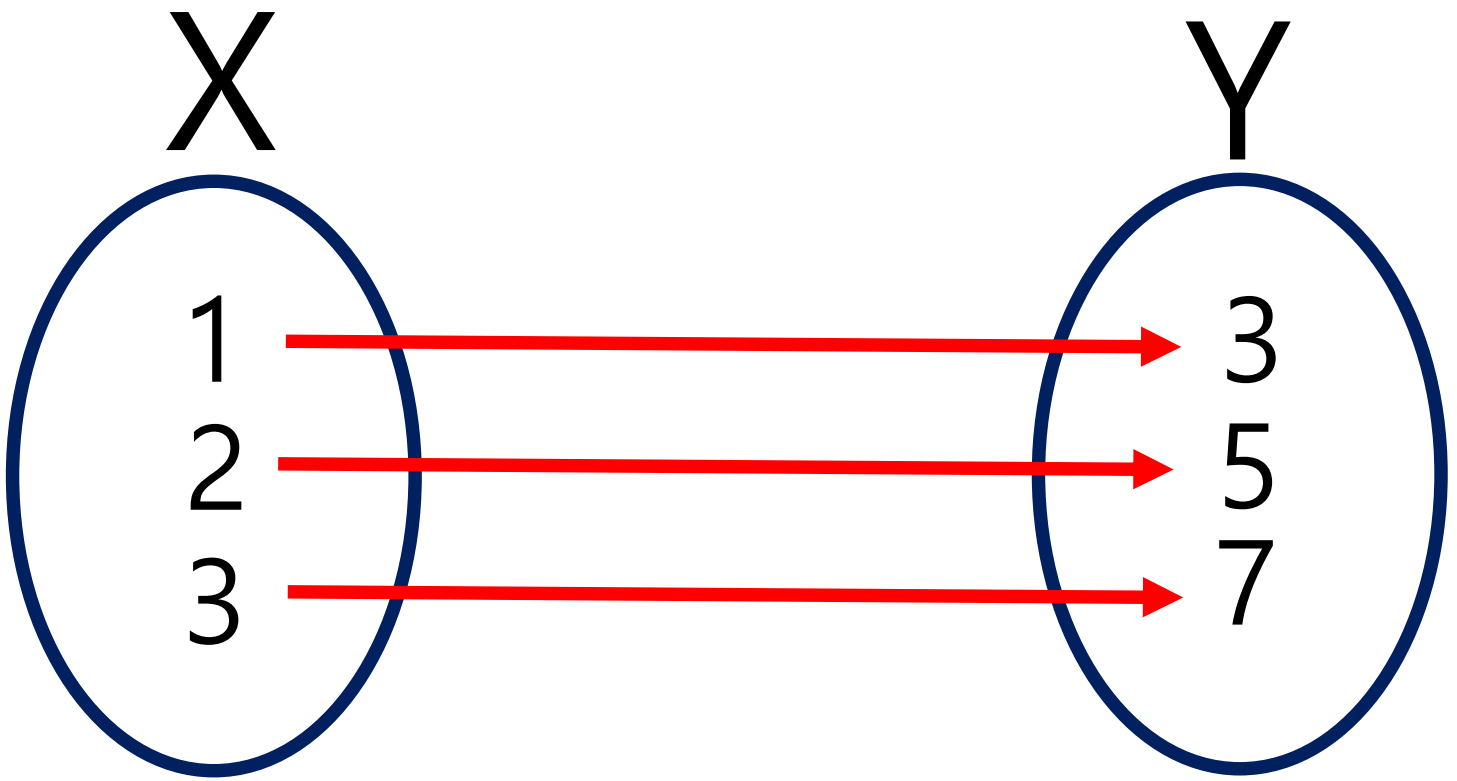
- 아버지의 키가 아들의 키에 어떤 영향을 줄까? 아버지의 키와 아들의 키 사이의 관계는?
- 고양이와 강아지를 어떻게 구분할까?
- “관계”, “연관성”을 수학적으로 표현하기

데이터사이의 관계 찾기 – 함수

- 아들의 키 = f(아버지의 키)+ 나머지
- f(고양이 이미지) -> 1, f(강아지 이미지) -> 0
- “f”를 어떻게 찾을까?

함수란?

$$f : X \rightarrow Y$$



함수관계에 있는 것들?

- 아버지와 아들의 키

- 아들의 키 = f(아버지의 키)+ 나머지
- 나머지: 어머니의 키, 발육상태, 집안환경, 관측오차

함수관계에 있는 것들?

- 강수량과 교통사고의 수

- 교통사고의 수 = f(강수량)+ 나머지
- 나머지: 교통사고 위치, 교통사고 시간, 자동차의 종류, 운전자 연령, 관측오차

함수관계에 있는 것들?

- 고양이 사진 분류기

- 고양이 여부 = f(사진) + 나머지
- 나머지: 사진의 크기, 해상도 등

함수관계에 있는 것들?

- 나이와 핸드폰 앱별 사용비율

- (유튜브, 인스타그램, 틱톡, 기타) = f(나이, 성별, 거주지역, 사용언어) + 나머지
- 나머지: 데이터플랜 등

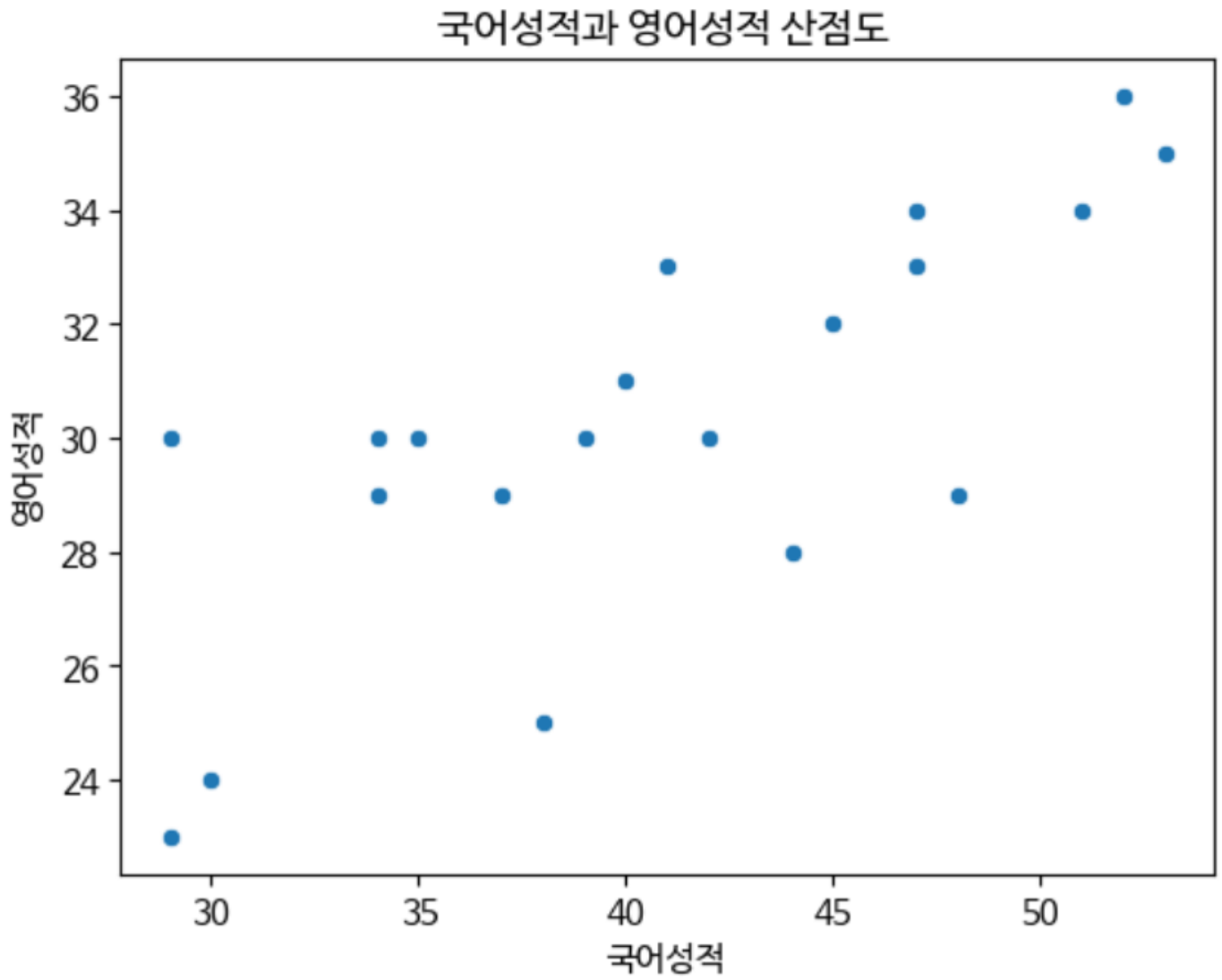
함수관계를 찾고 싶다?

- 국어점수와 영어점수
- 정의역? 공역?
- 연관성과 인과관계

국어점수, 영어점수 데이터

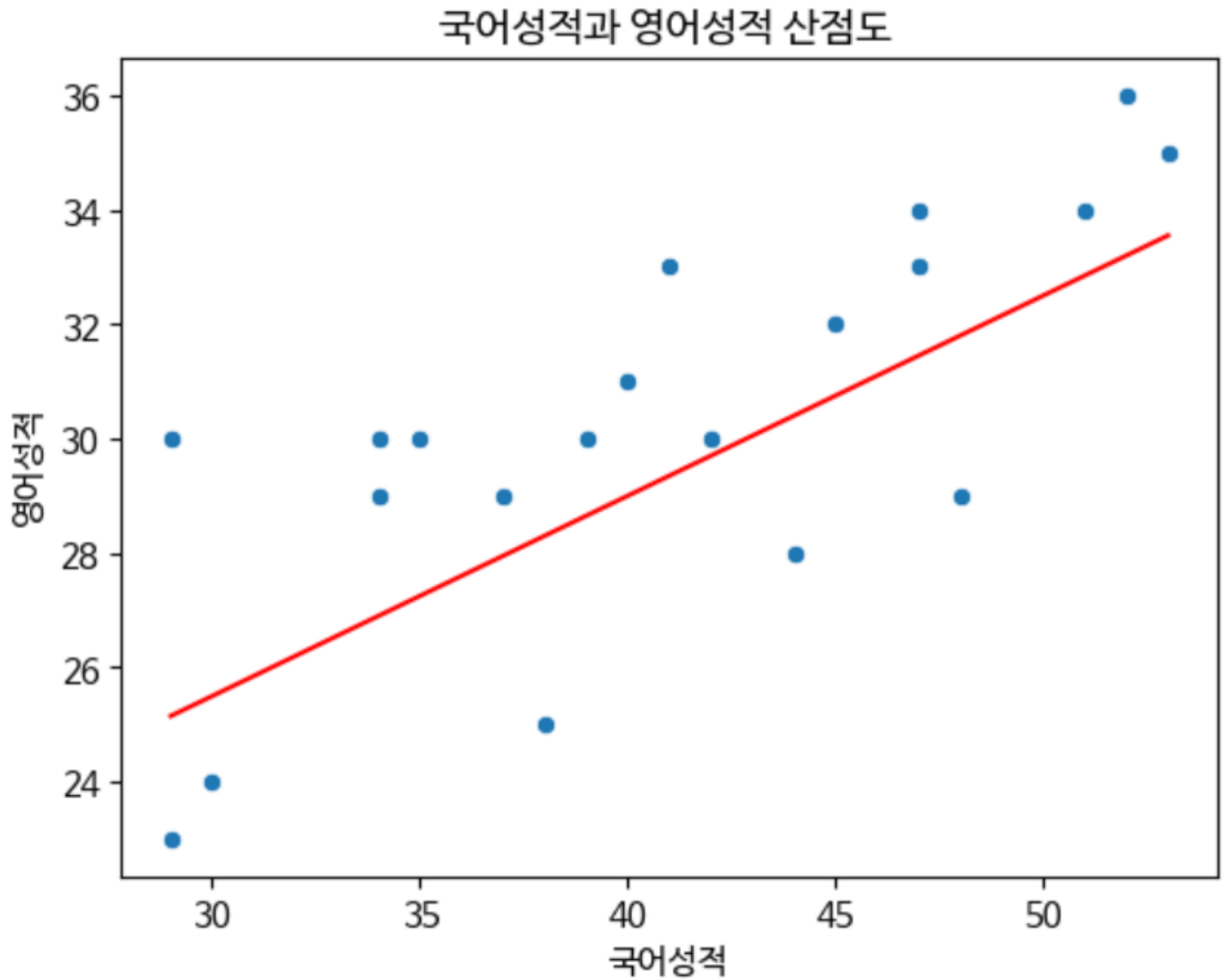
학생번호	1	2	3	4	5	6	7	8	9	10
국어	42	38	51	53	40	37	41	29	52	39
영어	30	25	34	35	31	29	33	23	36	30
학생번호	11	12	13	14	15	16	17	18	19	20
국어	45	34	47	35	44	48	47	30	29	34
영어	32	29	34	30	28	29	33	24	30	30

국어점수, 영어점수 산포도



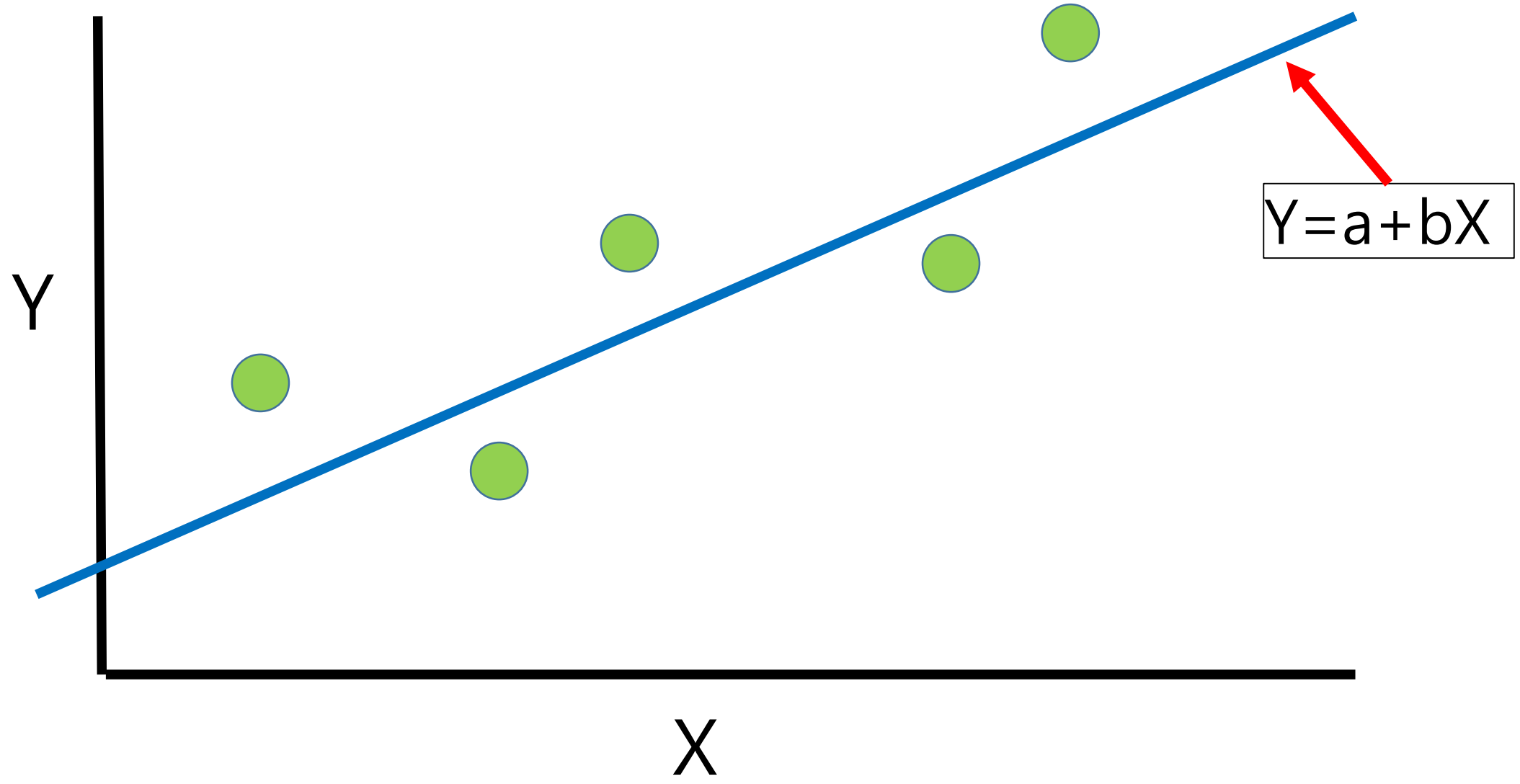
쉬운 것부터…

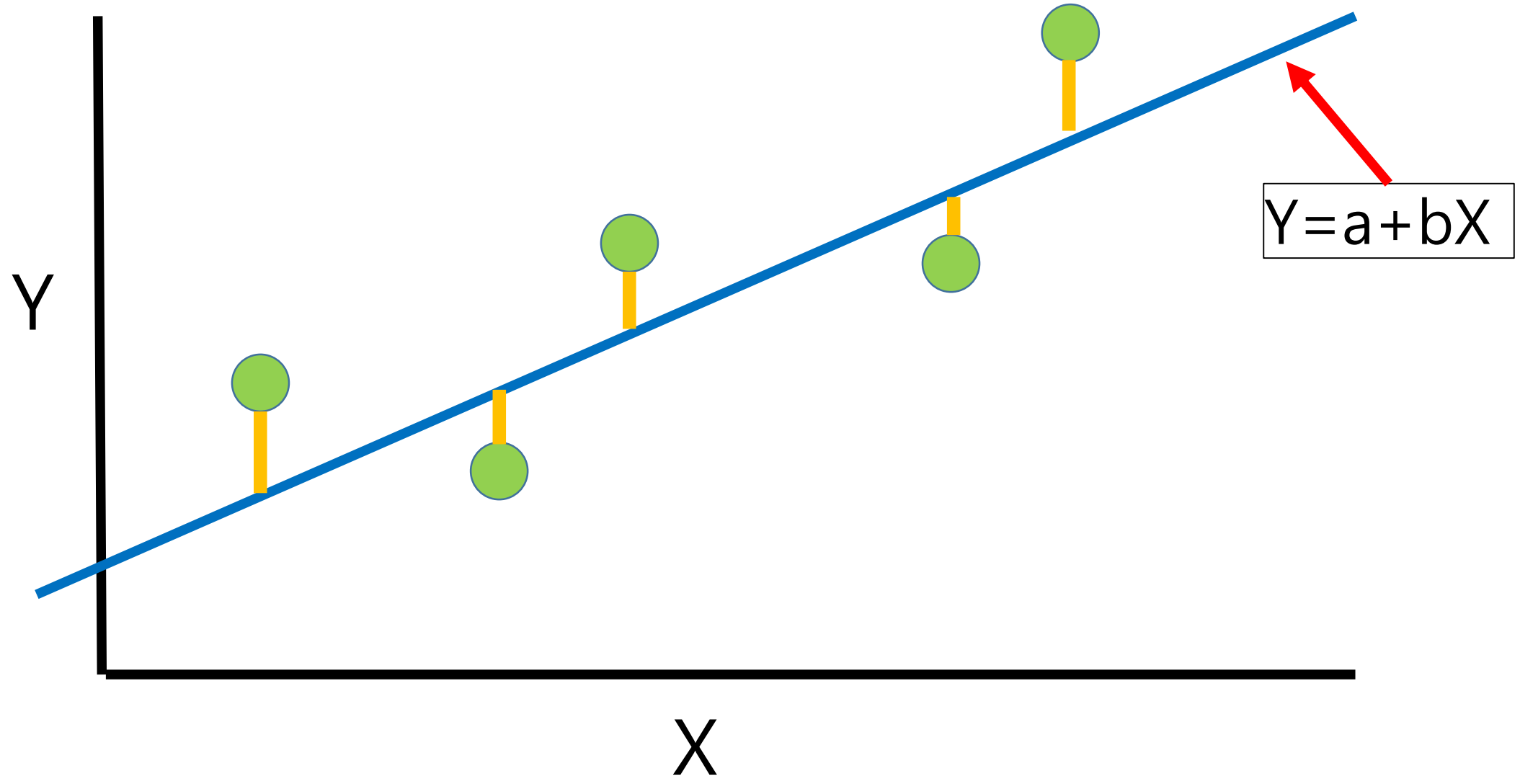
- 영어점수=f(국어점수)
- f(국어점수)=??


국어성적과 영어성적 산점도
영어성적
국어성적
36
34
32
30
28
26
24
30
35
40
45
50

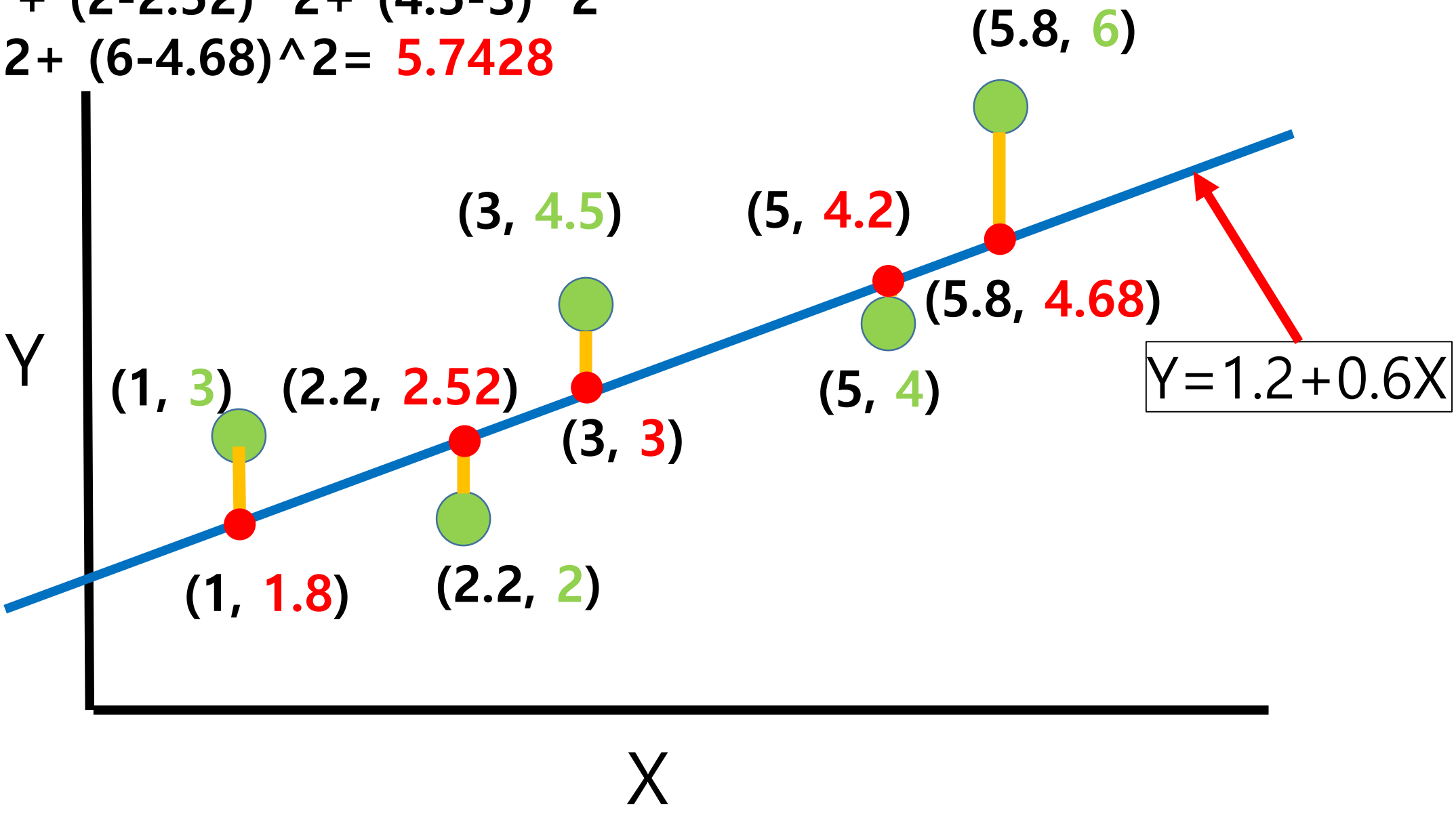
선형 모형

- 영어점수= a+ b × (국어 점수) + 오차
- $Y_i = a + bX_i + e_i$
- a, b를 어떻게 찾을까?

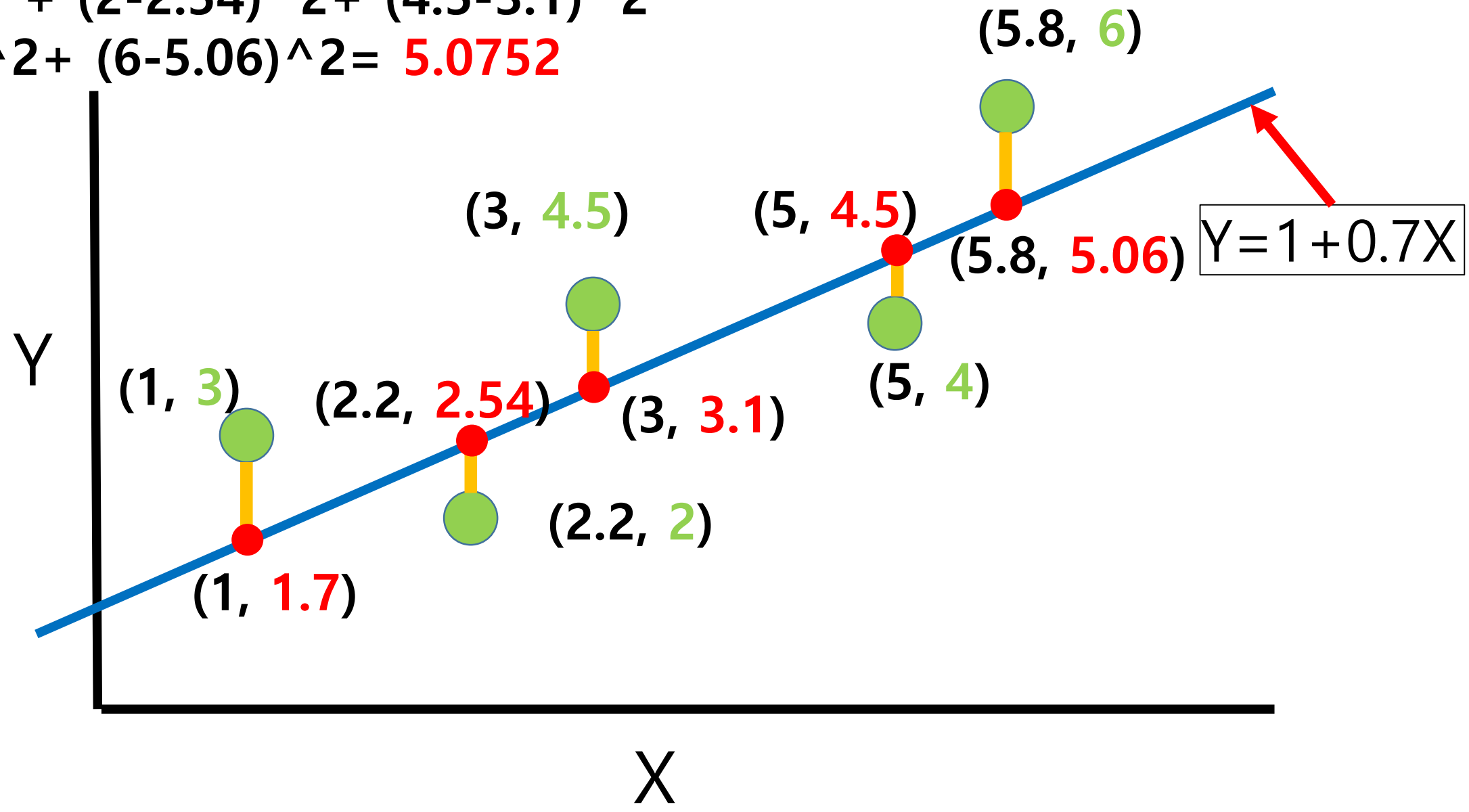

Y=a+bX
Y
X


Y=a+bX
Y
X

$(3-1.8)^2 + (2-2.52)^2 + (4.5-3)^2 + (4-4.2)^2 + (6-4.68)^2 = 5.7428$



(3-1.7)^2 + (2-2.54)^2+ (4.5-3.1)^2
+(4-4.5)^2+ (6-5.06)^2= 5.0752



최소 제곱법

오차 $= Y_i - (a + bX_i)$

오차제곱 $= (Y_i - (a + bX_i))^2$

오차제곱합 $= S(a,b) = \sum_{i=1}^{n} (Y_i - (a + bX_i))^2$

오차제곱합을 최소로 하는 a, b를 찾기

앞의 예시에서

데이터: (1,3), (2.2, 2), (3,4.5), (5,4), (5.8,6)

오차제곱합 $= S(a,b)=\sum_{i=1}^{n}(Y_i-(a+bX_i))^2 =$

$$(3-a-b\cdot 1)^2+(2-a-b\cdot 2.2)^2+(4.5-a-b\cdot 3)^2$$

$$+(4-a-b\cdot 5)^2+(6-a-b\cdot 5.8)^2$$

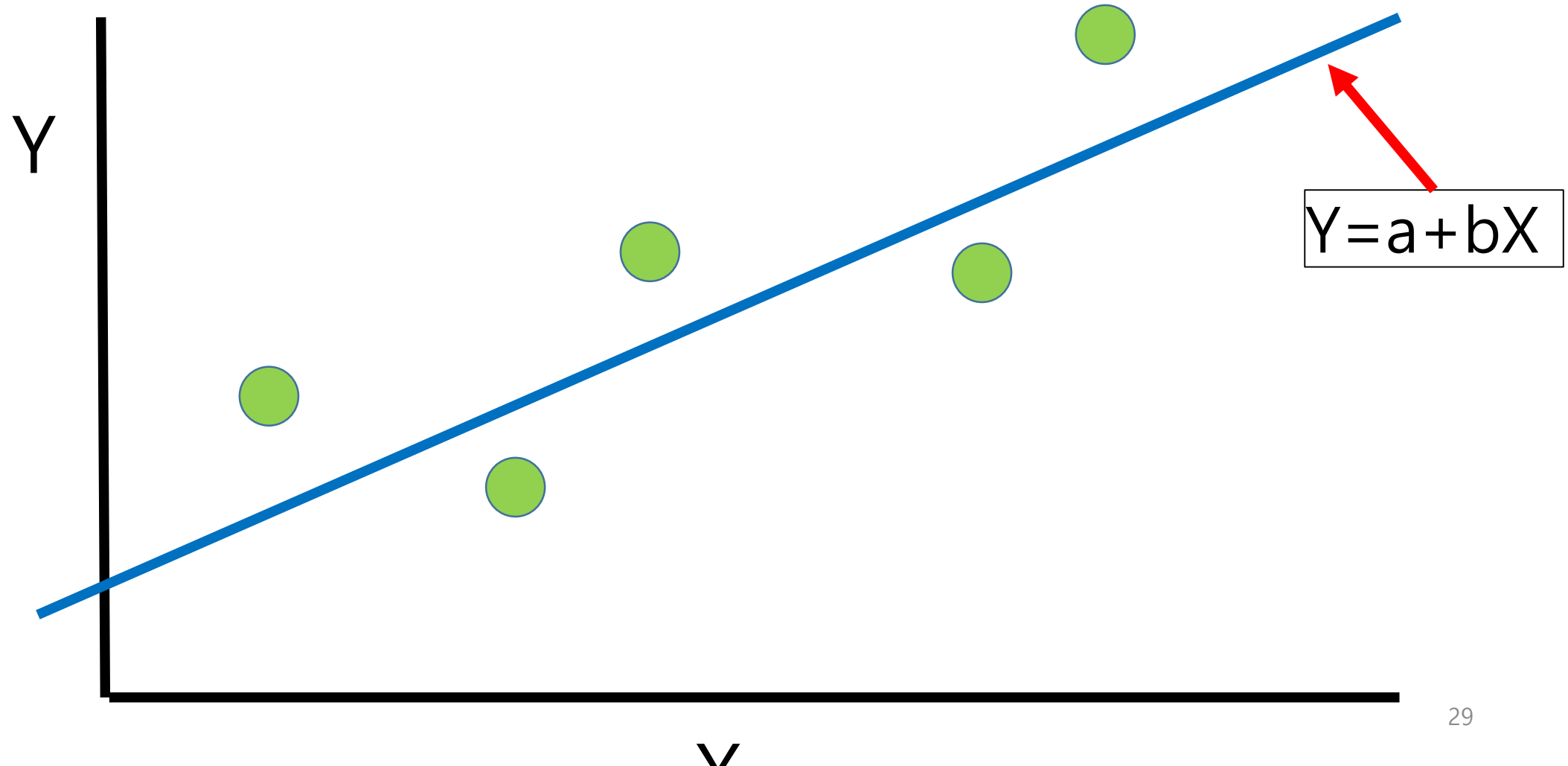
$$=5a^2+73.48b^2-39a\ -151.4b+34ab+85.25$$

- 오차제곱합을 최소로 하는 a, b를 찾기 (힌트: 2차 함수)
- 찾은 a, b를 $\hat{a}, \hat{b}$ 라고 하자.
- $\hat{a} = ?$
- $\hat{b} = ?$
- $S(\hat{a}, \hat{b}) = 3.5648$

선형회귀분석

- 선형모형을 가정하고 최소 제곱법으로 a, b를 찾고 결과를 해석하는 분석을 선형회귀분석이라고 한다.
- 데이터가 많아지면 컴퓨터로 a, b를 찾아야 한다.

다른 방법?



실제 데이터로 관계 찾기

국어/영어 성적 데이터를 이용해보기

학생번호	1	2	3	4	5	6	7	8	9	10
국어	42	38	51	53	40	37	41	29	52	39
영어	30	25	34	35	31	29	33	23	36	30
학생번호	11	12	13	14	15	16	17	18	19	20
국어	45	34	47	35	44	48	47	30	29	34
영어	32	29	34	30	28	29	33	24	30	30

- 오차제곱합을 최소로 하는 a, b를 찾기
- 오차제곱합 $= S(a,b) = \sum_{i=1}^{n} (Y_i - (a + bX_i))^2$

$$= (30 - a - b \cdot 42)^2 + (25 - a - b \cdot 38)^2$$

$$+ \ \ldots$$

- 컴퓨터로 해보자!

데이터분석 도구

- 다양한 데이터분석 도구: 대화형 언어
- R (www.r-project.org)
- 파이썬(www.python.org)

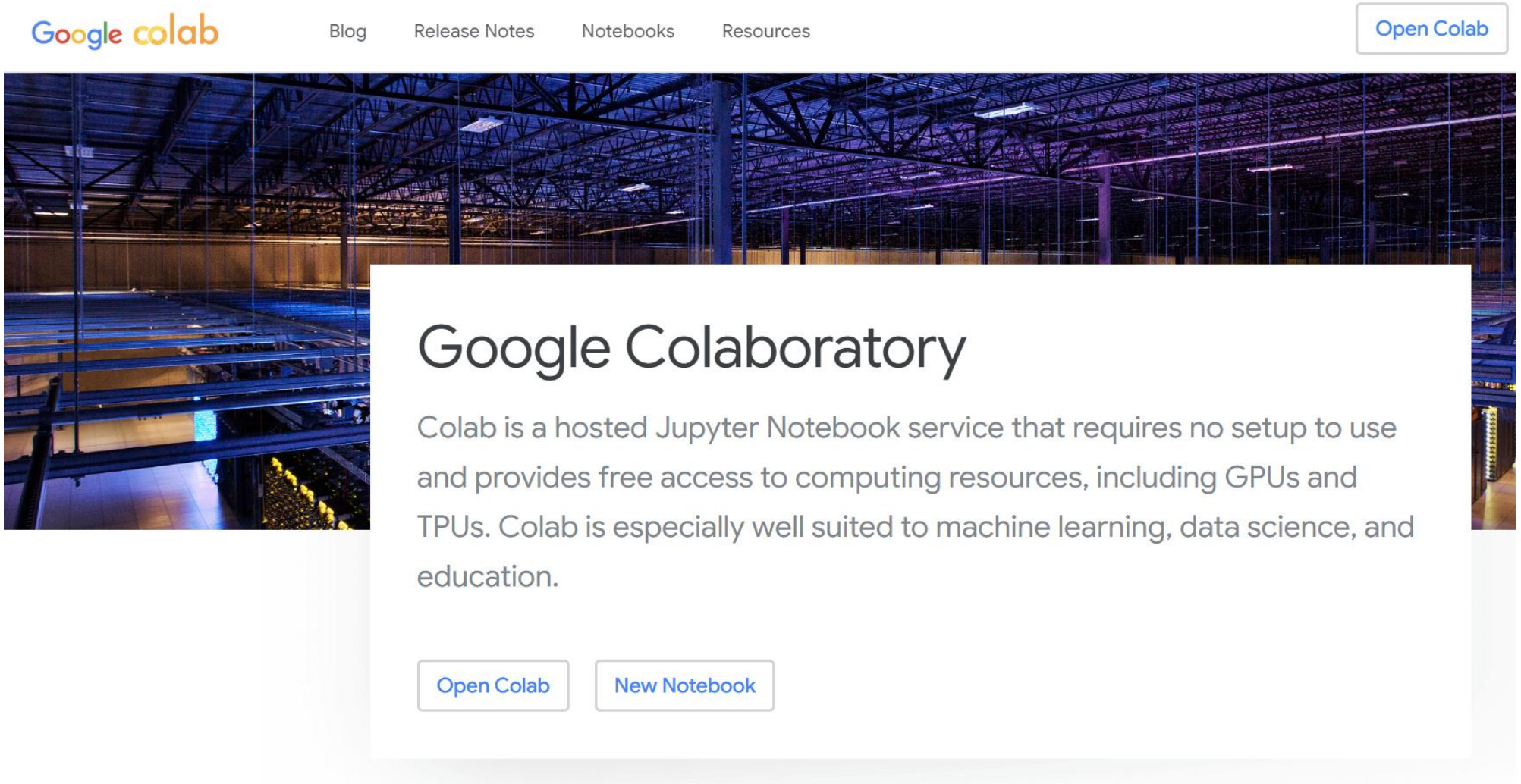
구글 코랩 사용하기

- Google Colaboratory의 줄임말.
- 클라우드 기반 Python을 작성하고 실행 가능한 무료 서비스.
- 별도의 Python설치 및 Module 설치 필요하지 않음.

구글 코랩 사용하기

- 구글계정 필요 하며 https://colab.research.google.com/notebooks/intro.ipynb?utm_source=scs-index로 접속 또는 구글에서 Google Colab 검색
- 구글 코랩에서 텍스트는 마크다운(Markdown) 언어 사용

구글 코랩 사용하기


Google colab
Blog
Release Notes
Notebooks
Resources
Open Colab
Google Colaboratory
Colab is a hosted Jupyter Notebook service that requires no setup to use and provides free access to computing resources, including GPUs and TPUs. Colab is especially well suited to machine learning, data science, and education.
Open Colab
New Notebook

국어/영어 성적 데이터를 입력

[20]
```
# 데이터를 저장, 처리하기위한 pandas 라이브러리 불러오기
import pandas as pd
# 숫자 연산을 효율적으로 하기 위한 라이브러리 불러오기
import numpy as np
```

[10]
```
# pandas 라이브러리의 데이터프레임 (DataFrame)을 사용하여 작은 크기의 데이터 직접입력
score = pd.DataFrame(
    { 'kor' : [42, 38, 51, 53, 40, 37, 41, 29, 52, 39, 45, 34, 47, 35, 44, 48, 47, 30, 29, 34],
      'eng'  : [30, 25, 34, 35, 31, 29, 33, 23, 36, 30, 32, 29, 34, 30, 28, 29, 33, 24, 30, 30]
    } )
score
```

선형회귀모형 적합:

```
score_fit = smf.ols(formula='eng ~ kor', data=score).fit()
score_fit.params
```

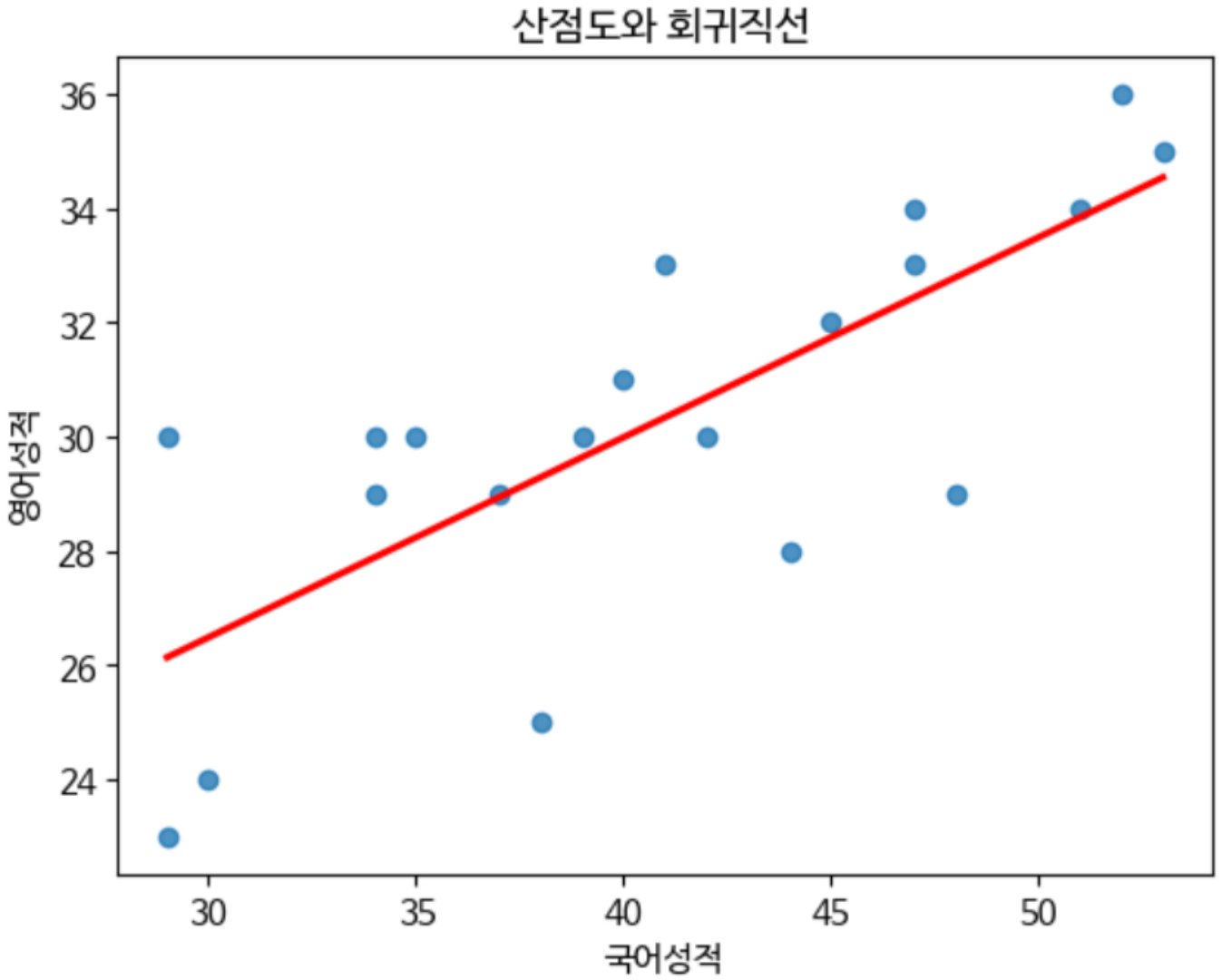
	0
Intercept	15.989850
kor	0.349942

dtype: float64

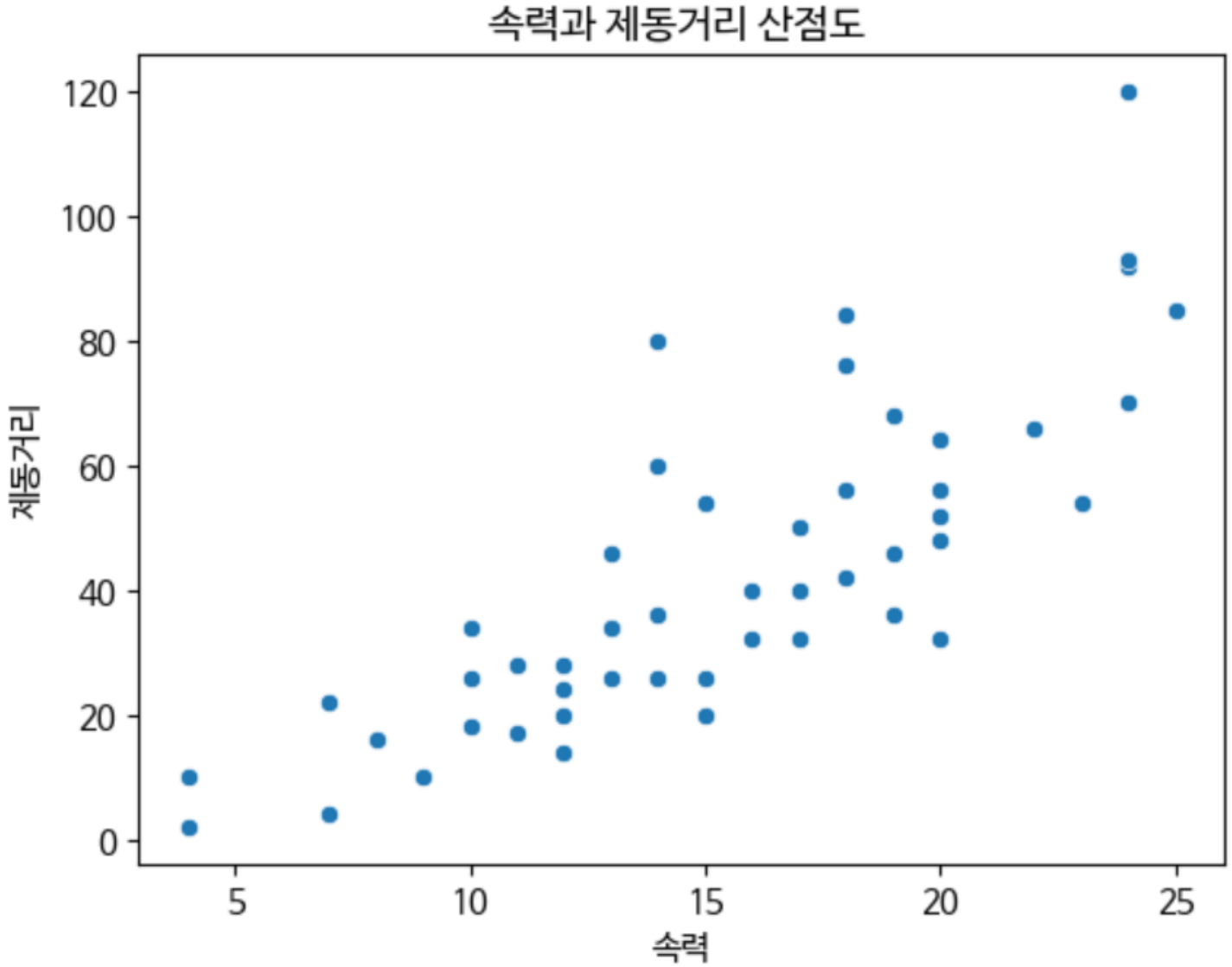
결과 해석하기

- 영어점수$= 15.9899+0.3499 \times$ 국어점수
- $Y = 15.9899 + 0.3499\ X$
- 0.3499 해석
- 15.9899 해석

회귀직선 같이 그리기



예제: 속력과 제동거리



선형회귀모형 적합

```
[38] cars_fit = smf.ols(formula='dist ~ speed', data=cars).fit()
     cars_fit.params
```

	0
Intercept	-17.579095
speed	3.932409

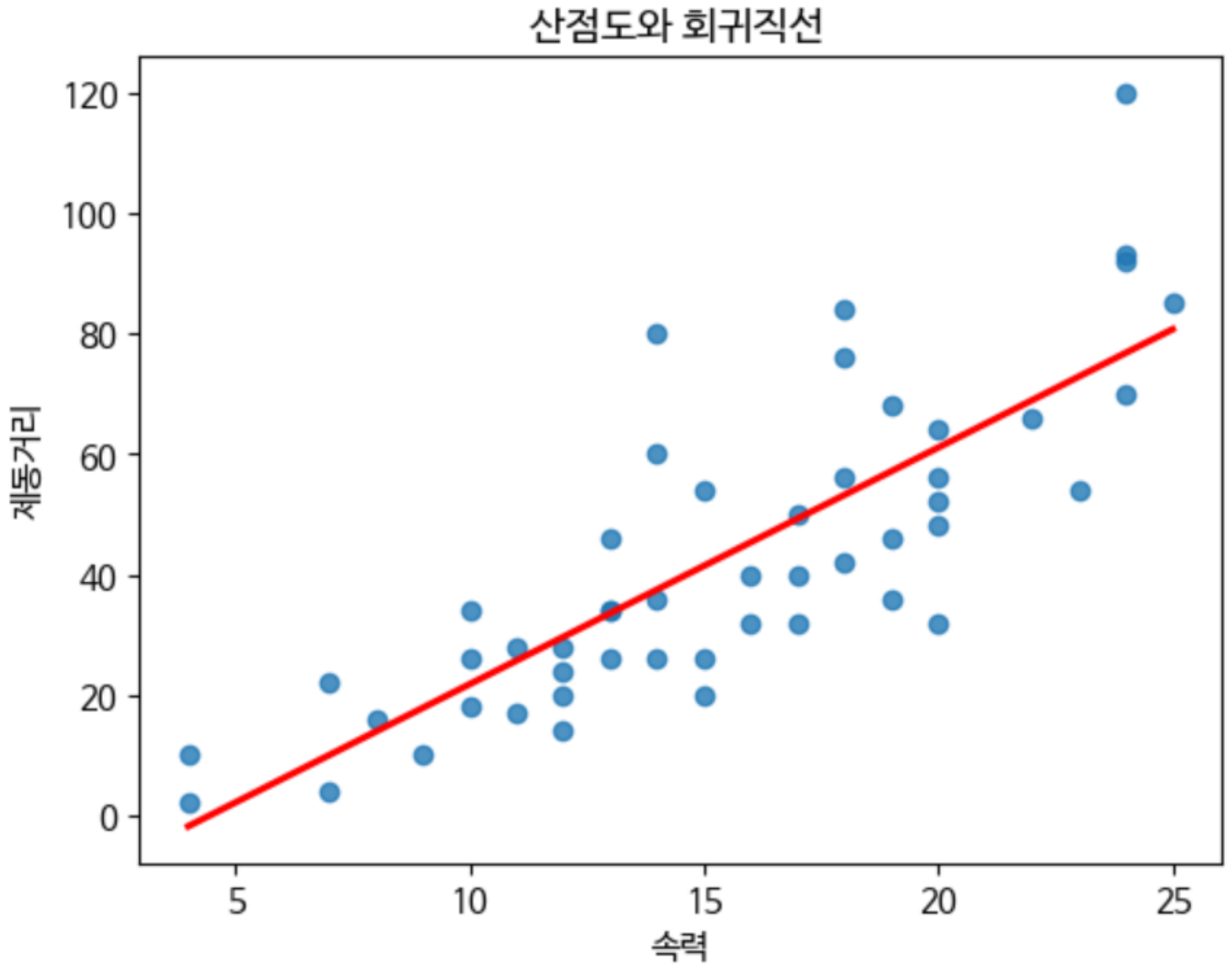
dtype: float64

해석하기

제동거리(ft)= -17.579+3.932 ×속력(mph)

$Y = -17.579 + 3.932\ X$

회귀직선 같이 그리기



얼마나 잘 설명하나?

- 선형모형: $Y = a + b\,X + e$
- 추정된 회귀직선: $Y = -17.579 + 3.932\,X$
- 추정된 오차 (잔차):
  - 데이터와 추정된 모형(추정된 회귀직선)사이의 차이
- $\hat{e}_i = Y_i - \hat{a} - \hat{b}\,X_i$

얼마나 잘 설명하나?

- 잔차제곱합: $\mathrm{SSE} = \sum_{i=1}^{n}(Y_i - \hat{a} - \hat{b}\,X_i)^2$
- 총 제곱합: $SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$
- 회귀제곱합: $SSR = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, $\hat{Y}_i = \hat{a} + \hat{b}\,X_i$
- $R^2 = SSR/SST$: 회귀모형으로 설명할 수 있는 Y의 변동성 : 결정계수

각 제곱합과 결정계수 구해보기

[65]
```
# 잔차제곱합 계산하기
SSE= (cars_fit.resid**2).sum()
SSE
```

```
np.float64(11353.52105109489)
```

[66]
```
# 총 제곱합 계산하기
y_mean = cars['dist'].mean()
SST = ((cars['dist'] - y_mean)**2).sum()
SST
```

```
np.float64(32538.980000000003)
```

[70]
```
# 결정계수 구하기
Rsquared = 1 - SSE/SST
Rsquared
```

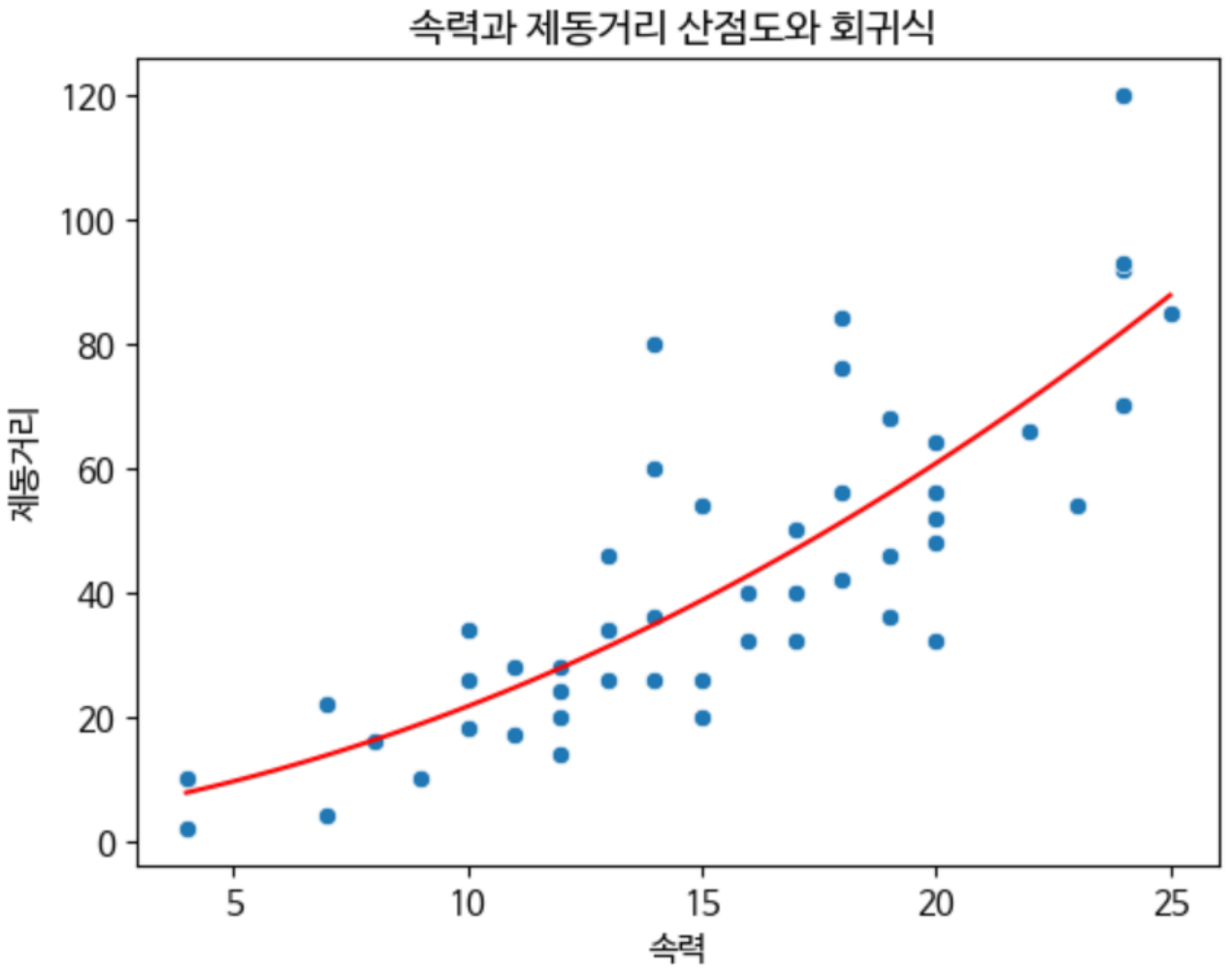
```
np.float64(0.6510793807582509)
```

결정계수를 높이기

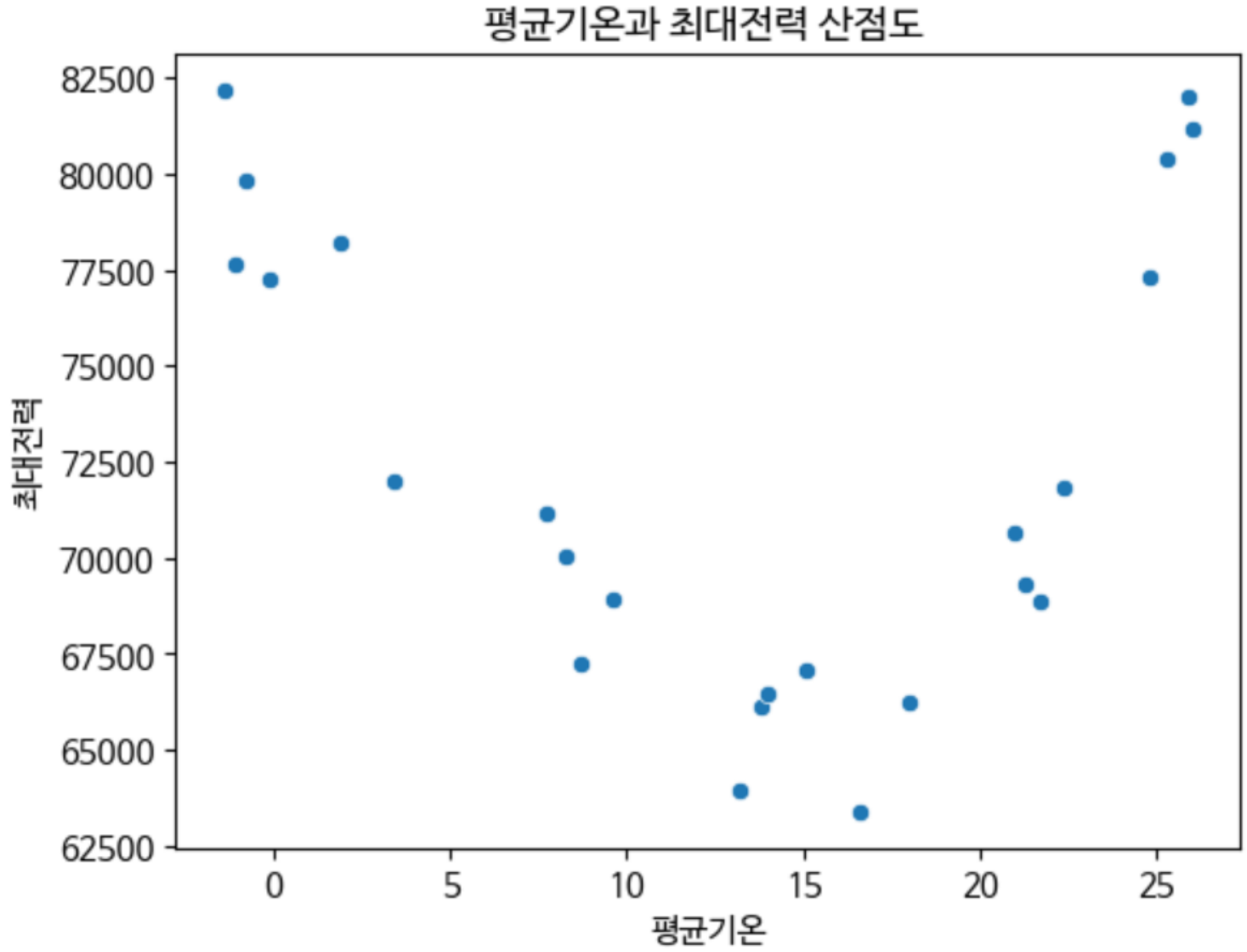
- 제곱항 추가
- $Y_i = a + b\,X_i + cX_i^2 + e$

```
cars_fit2.rsquared
```

```
np.float64(0.6673308165262097)
```



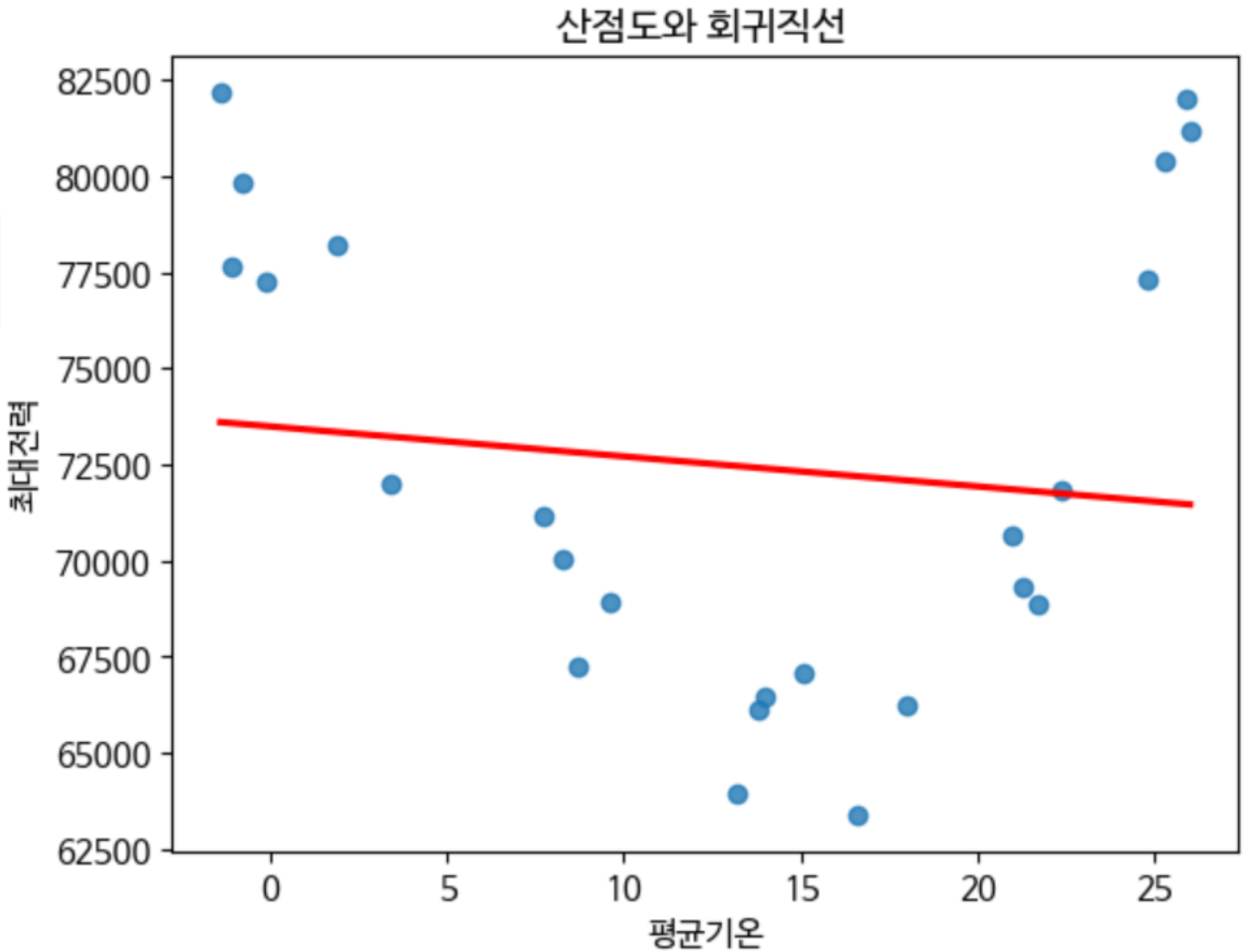
예제: 평균기온과 최대전력



회귀직선 같이 그리기

```
[81] power_fit.rsquared
```

```
np.float64(0.014431093123215444)
```

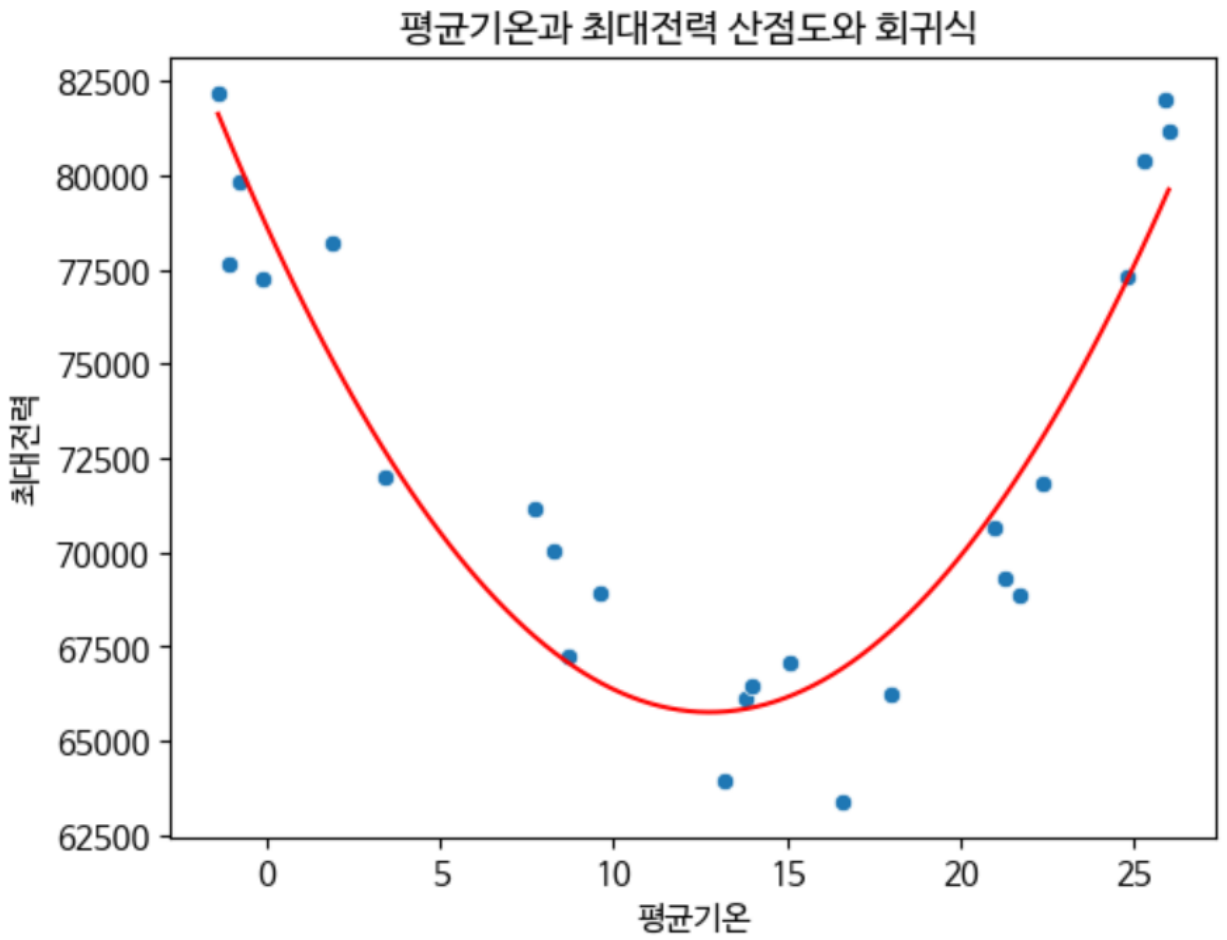


결정계수를 높이기

- 제곱항 추가
- $Y_i = a + b\,X_i + cX_i^2 + e$

```
[83]  power_fit2.rsquared
```

```
np.float64(0.8841830828948174)
```



함수관계 찾기에서 조심할 것들..

- 네 개의 데이터셋을 보자.
- X의 평균
- Y의 평균
- X의 분산
- Y의 분산

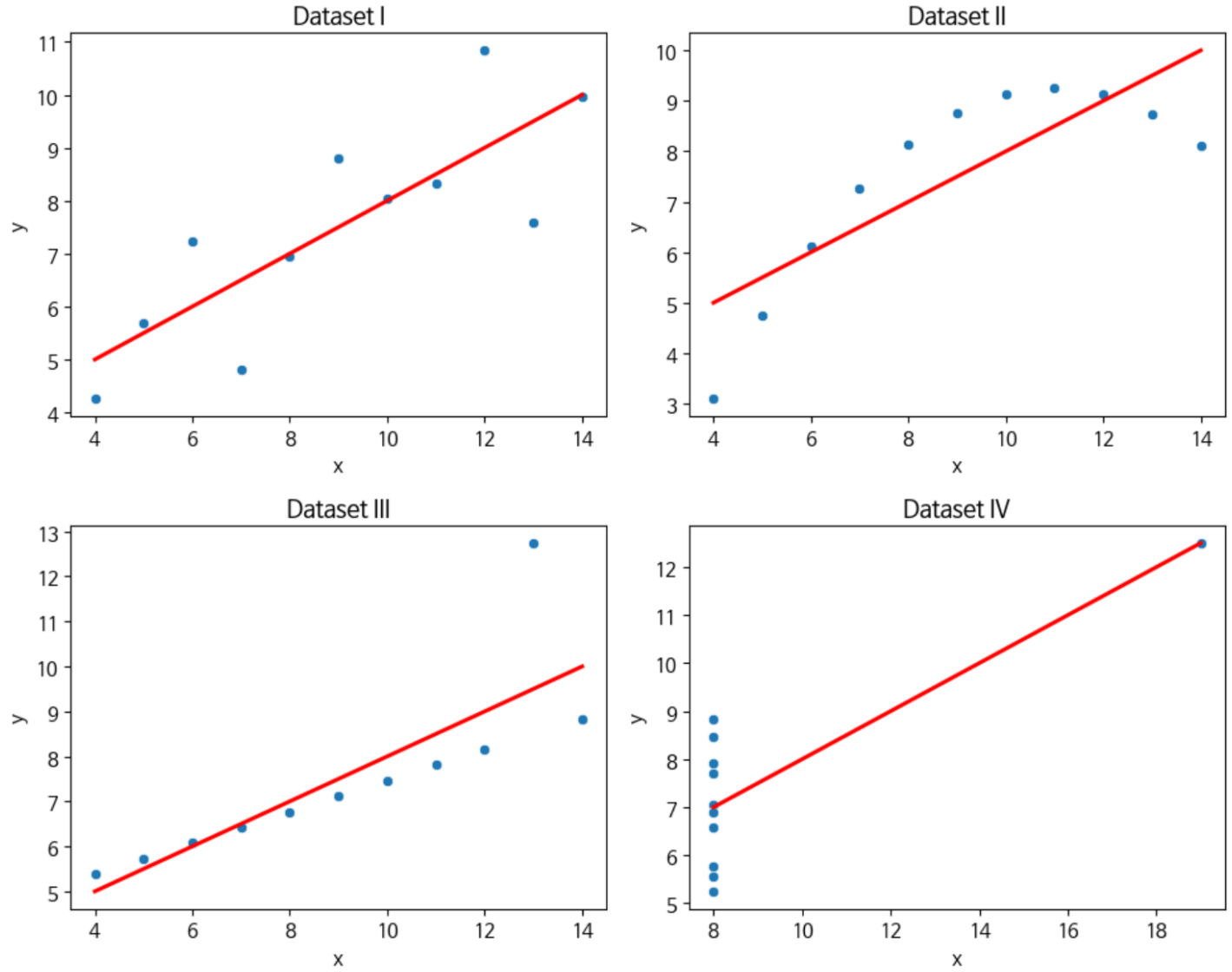
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Anscombe’ quartet

```
Dataset: I
  Mean of x: 9.0
  Variance of x: 11.0
  Mean of y: 7.500909090909093
  Variance of y: 4.127269090909091
  Correlation between x and y: 0.816420516344 84
--------------------
Dataset: II
  Mean of x: 9.0
  Variance of x: 11.0
  Mean of y: 7.50090909090909
  Variance of y: 4.127629090909091
  Correlation between x and y: 0.8162365060002428
```

```
Dataset: III
  Mean of x: 9.0
  Variance of x: 11.0
  Mean of y: 7.5
  Variance of y: 4.12262
  Correlation between x and y: 0.8162867394895984
--------------------
Dataset: IV
  Mean of x: 9.0
  Variance of x: 11.0
  Mean of y: 7.500909090909091
  Variance of y: 4.123249090909 09
  Correlation between x and y: 0.8165214368885028
```

Anscombe’ quartet



데이터의 변동성 반영하기

데이터의 변동성이란?

- 데이터를 추출할때마다 다른 데이터가 관측될 수 있음

예) 동전던지기 10번을 할 때

- 데이터의 변동성에 대한 수학적 설명: 예) 확률변수의 분산
- 분석한 결과도 변동성을 가짐
- 변동성이 작은지/큰지에 따라 분석한 결과의 신뢰도가 영향을 받음

선형모형의 추정된 b(기울기)의 변동성

- $Y=a+bX+$ 나머지
- $b=0$이면?
- $\hat{b}=0.1$ 이면 0이라고 할 수 있을까?
- 데이터의 변동성을 통해 추정된 b의 변동성을 알게 되면

$\hat{b}=0.1$ 는 $b=0$이었을 때 충분히 일어날 수 있는 일인지 드물게 일어나는 일인지 판단 할 수 있게 됨.

(속력,제동거리) 데이터

```
#속력, 제동거리 데이터에서 처음 25개로 회귀모형 적합하기
cars3 = cars.head(25).drop(columns=['speed_squared', 'residual'], errors='ignore')
cars3_fit = smf.ols(formula='dist ~ speed', data=cars3).fit()
cars3_fit.params
```

	0
Intercept	-10.003079
speed	3.289087

```
# 속력, 제동거리 데이터에서 25번째부터 50번째 데이터로 회귀모형 적합하기
cars4 = cars[25:50].drop(columns=['speed_squared', 'residual'], errors='ignore')
cars4_fit = smf.ols(formula='dist ~ speed', data=cars4).fit()
cars4_fit.params
```

	0
Intercept	-42.036446
speed	5.149921

Resampling (재표본화)

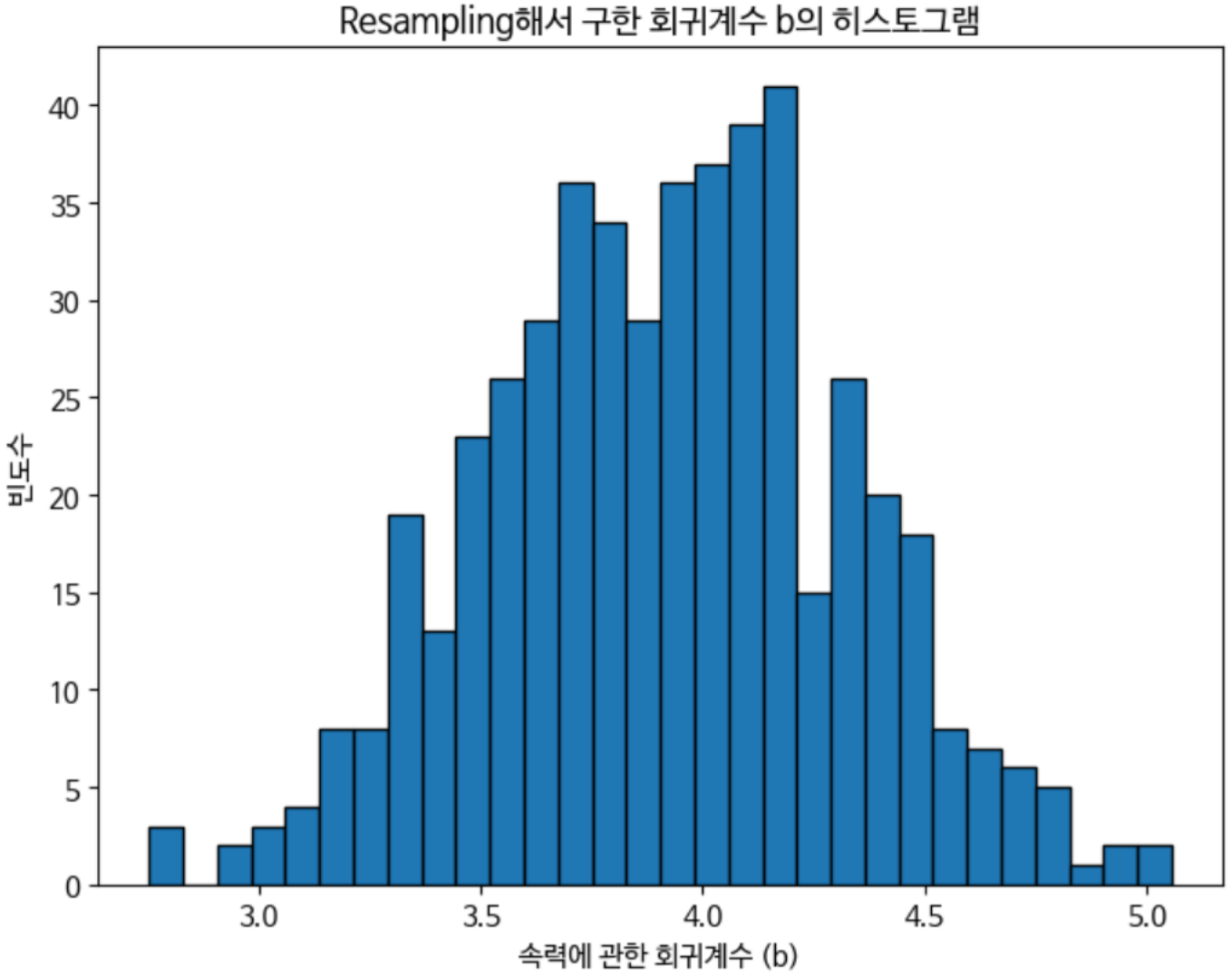
- 데이터가 한 세트 밖에 없으므로 데이터의 변동성을 흉내 내기위해 재표본화 (Resampling)을 함
- 가지고 있는 데이터 세트에서 다시 랜덤 추출(Resampling)해서 새로운 데이터 세트를 만들기를 반복
- 새로운 데이터 세트 마다 선형모형을 추정

Resampling (재표본화)

- 데이터세트마다 매번 다른 추정 값 $\hat{b}$이 나옴.
- 반복하면, 여러 개의 $\hat{b}$를 구할 수 있음. $\hat{b}_1, \hat{b}_2, \ldots$
- $\hat{b}_1, \hat{b}_2, \ldots$ 를 통해 변동성을 알 수 있음.

Resampling (재표본화)

- $\hat{b}$의 히스토그램



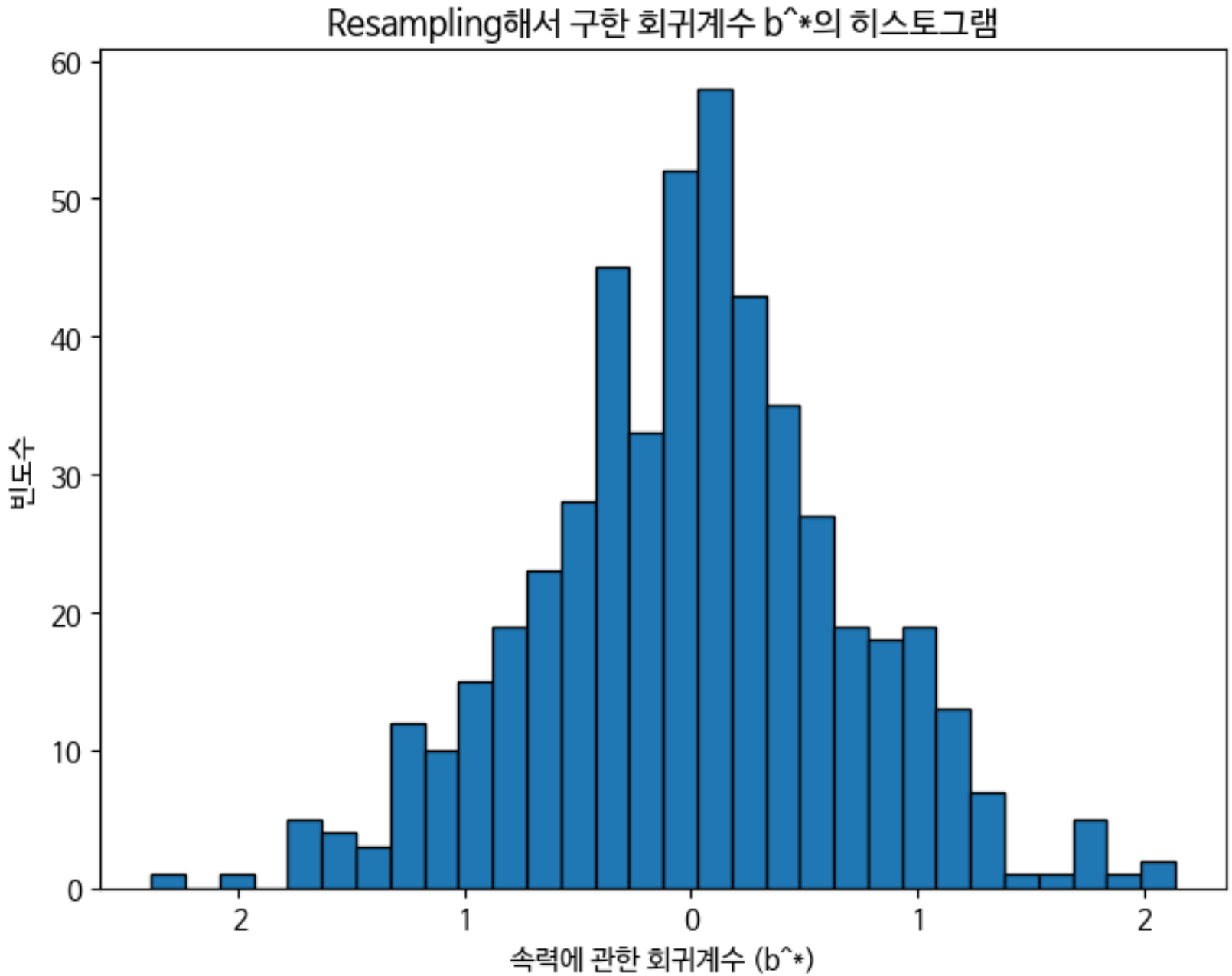
속력은 제동거리에 영향을 미치나?

- 제동거리$=$ a$+$ b 속력 $+$ 오차
- $\hat{a} =$ -17.579
- $\hat{b} = 3.932$
- b는 0이 아니라고 할 수 있나? 즉, 속력은 제동거리에 영향을 미치나?

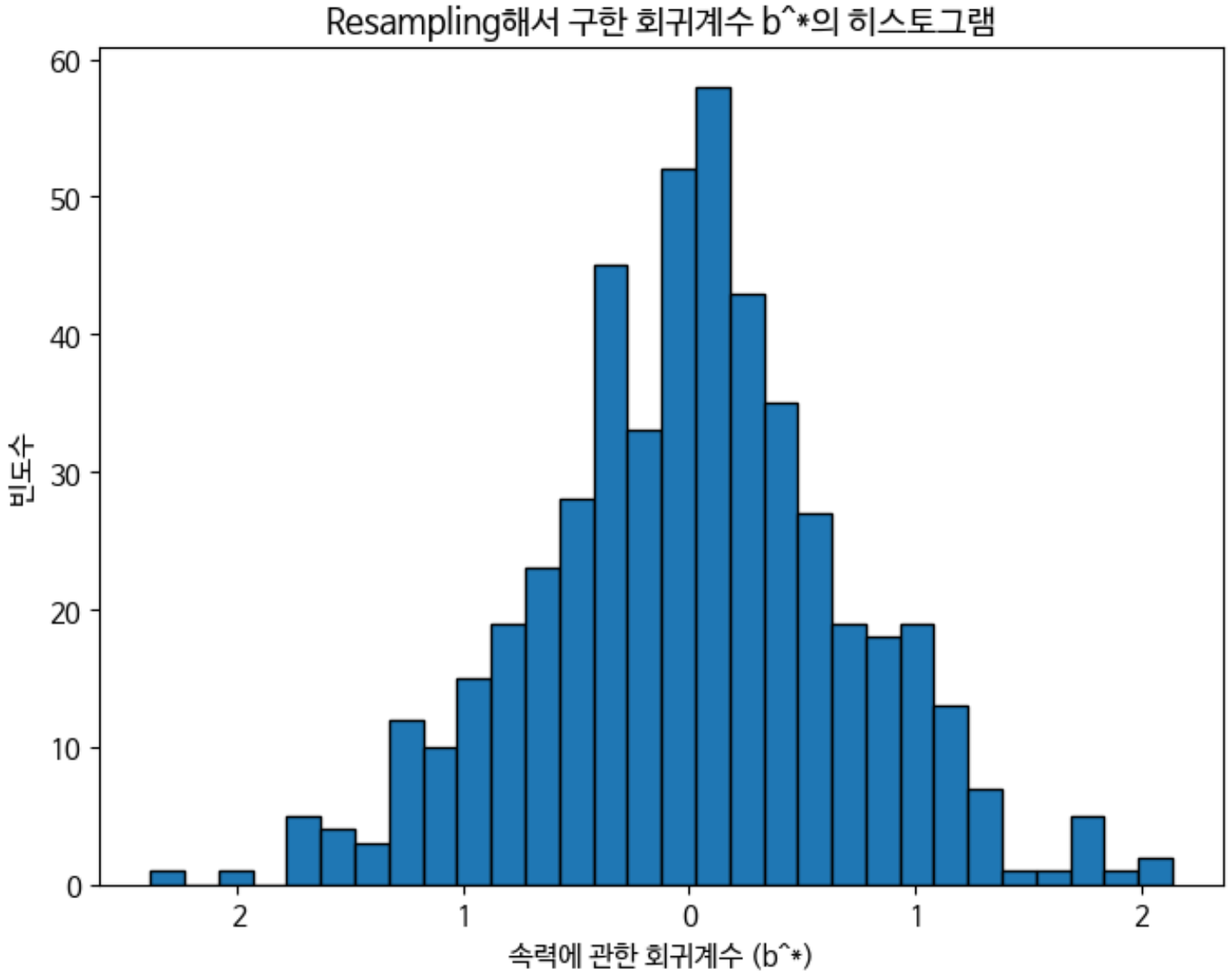
- b=0이 맞다면,
- 제동거리= a+ 0 속력+ 오차
- X=속력, Y=제동거리
- 관측한 데이터는 쌍으로 관측된 데이터 (Y_i, X_i)
- 제동거리와 속력사이에 아무 관계가 없다면, (Y_i, X_j) 여도 상관없음

- 순서가 섞인 쌍들로 이루어진 데이터 (Y_i, X_j) 로 선형모형을 적합하면 기존의 (Y_i, X_i)의 관계가 깨짐
- 즉, X, Y의 관계가 없어짐.
- 순서가 섞인 쌍들로 이루어진 데이터를 이용한 b의 추정값은 X,Y 사이의 관계가 없다고 가정했을 때의 b의 추정값이 됨.

- 순서가 섞인 쌍들로 이루어진 데이터 세트를 반복해서 만들고, 선형모형을 적합하면, 그때마다 b의 추정값 $\hat{b}^*$이 나옴.
- 이때 나온 $\hat{b}_1^*, \hat{b}_2^*, \ldots$ 들은 X,Y사이의 관계가 없다고 가정했을 때의 추정값들.
- 이 값들을 모아서 만든 분포는 "내가 가지고 있는 데이터 X,Y가 관계가 없다"고 했을 때의 분포.


Resampling해서 구한 회귀계수 b^*의 히스토그램
빈도수
60
50
40
30
20
10
0
2
1
0
1
2
속력에 관한 회귀계수 (b^*)

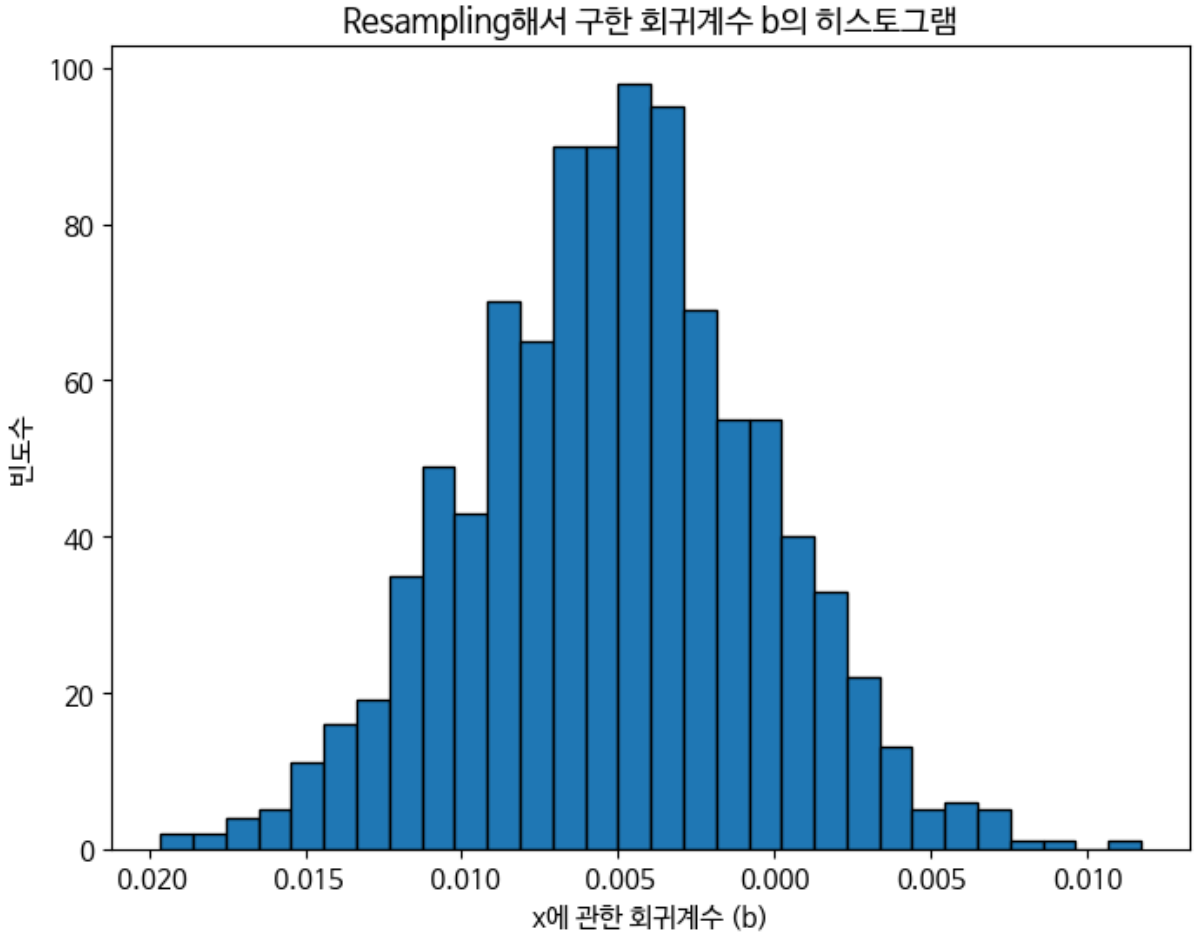
- 원래 데이터로 구한 $\hat{b}=3.932$는 어디쯤?



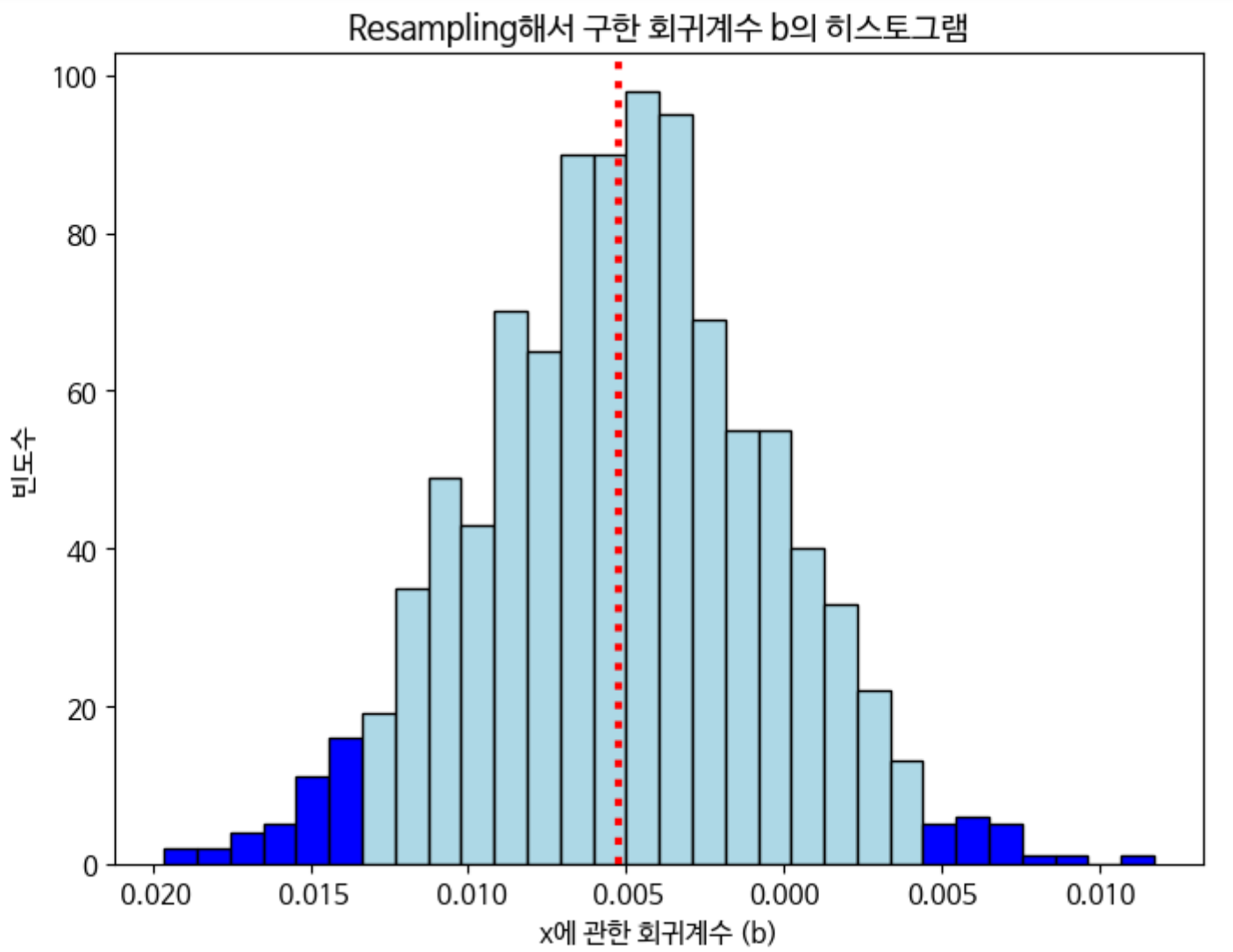
가상 데이터 예제 1

- 지역별 범죄 건수 (Y) = a + b 지역별 유흥업소 수 (X) + 오차
- 지역별 범죄건수는 월 평균이라고 하자.
- a= 50, b=0, X는 포아송 분포, 오차는 정규분포를 이용하여 가상 데이터 (Y_i, X_i)를 100개 생성한 후 b를 추정
- 이 경우 b=0이므로 X, Y는 서로 관련이 없음.

- 가상데이터에서 순서가 섞인 쌍들로 이루어진 데이터 세트를 반복해서 만들고, 선형모형을 적합해서, 그때마다 b의 추정 값 $\hat{b}^*$를 구해서 히스토그램을 그려봄.



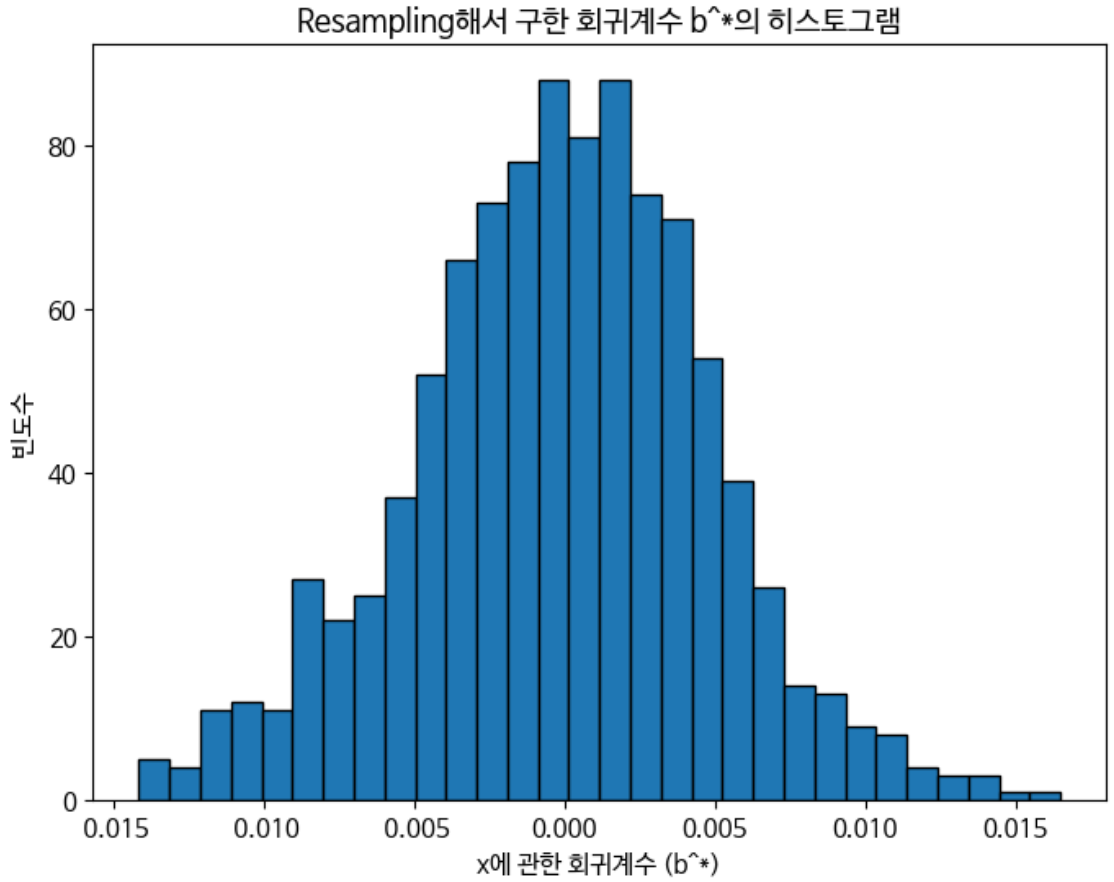
- 가상데이터 자체로 구한 $\hat{b}$=-0.005223 과 비교



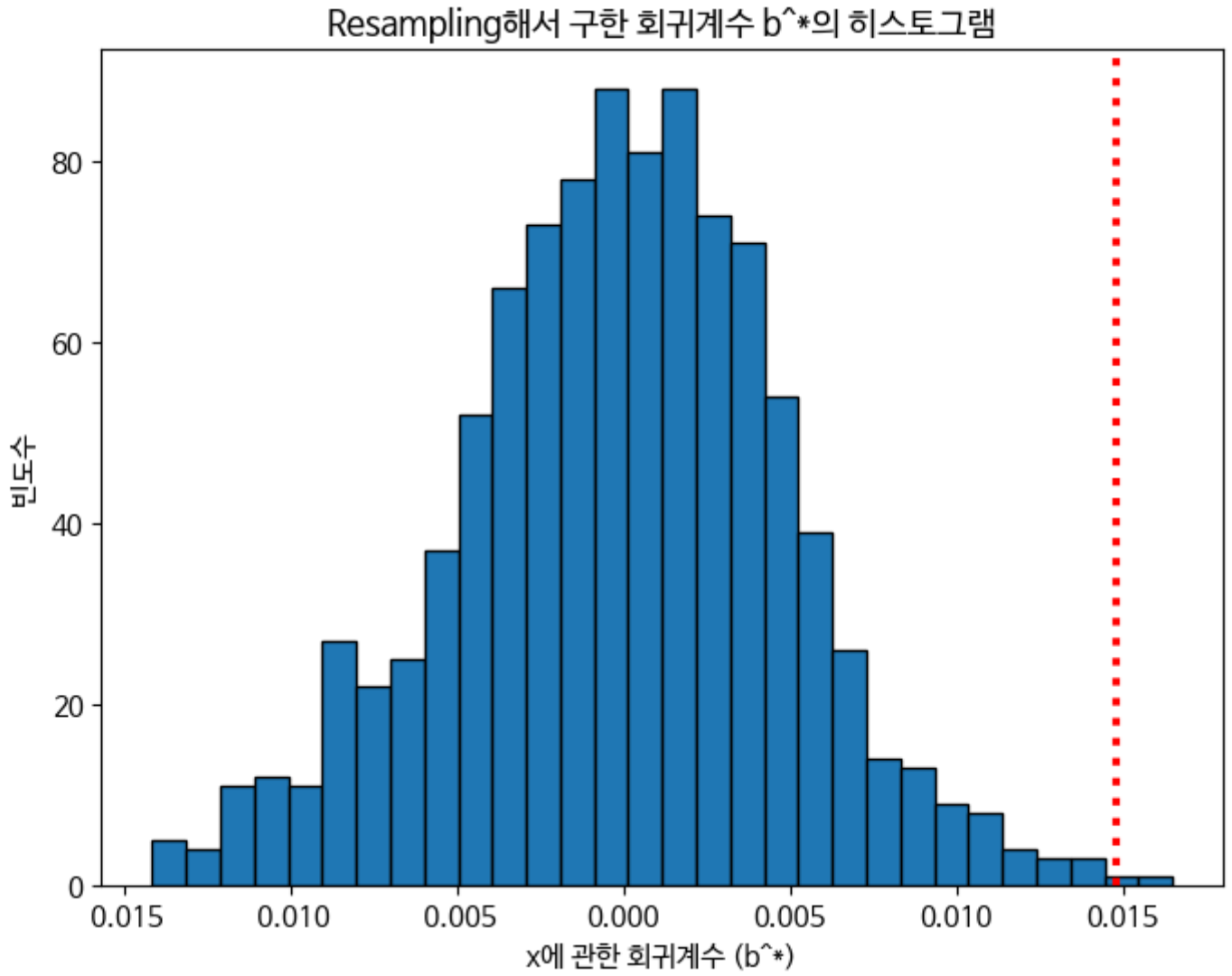
가상 데이터 예제 2

- 지역별 범죄 건수 (Y) = a + b 지역별 유흥업소 수 (X) + 오차
- a= 50, **b=0.02,** X는 포아송 분포, 오차는 정규분포를 이용하여 가상 데이터 (Y_i, X_i)를 100개 생성한 후 b를 추정
- 이 경우 b=0.02이므로 X, Y는 서로 관련이 있음.

- 가상데이터 2에서 x (지역별 유흥업소 수) 와 y (범죄 발생건수) 가 관계가 없다고 가정할때, 즉, x의 순서를 섞어서 만든 데이터 세트를 여러 개 만들어 추정한 회귀계수들의 히스토그램을 구해봄



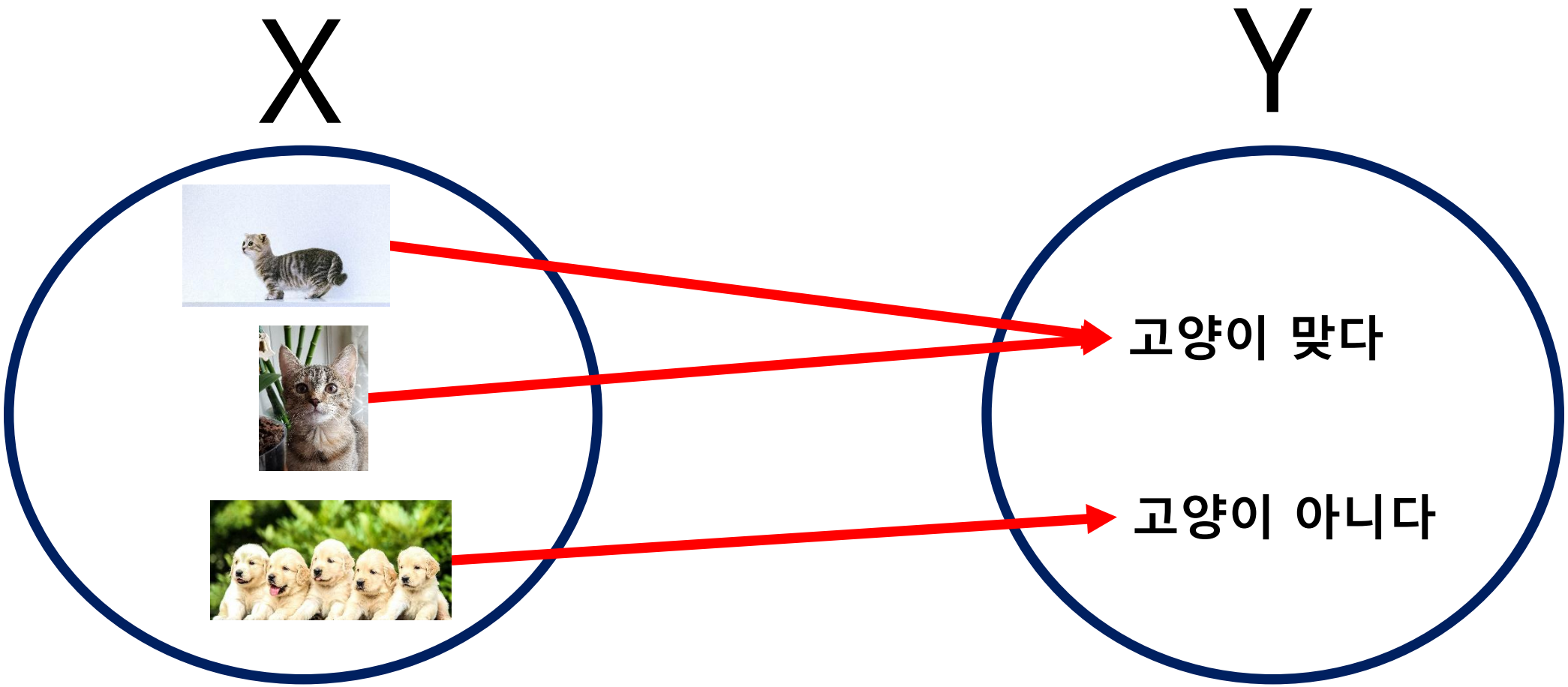
- 가상데이터 자체로 구한 $\hat{b}=0.014777$ 과 비교



분류 함수 찾기

Y가 연속적인 숫자가 아닌 경우의 함수관계

- 예: 고양이 사진 분류기



- Y는?
- 고양이 맞다 → 1, 고양이 아니다 → 0
- X는?
- f는?

데이터 예제

- 타이타닉 데이터
- 1912년 타이타닉호 침몰 사고 당시 승객들의 생존 여부(Survived)와 다양한 개인 정보(age, sex, class, fare 등)로 구성
- 이진 분류 문제 (생존 1 / 사망 0)

```python
# 데이터 확인
print(titanic.head())
```

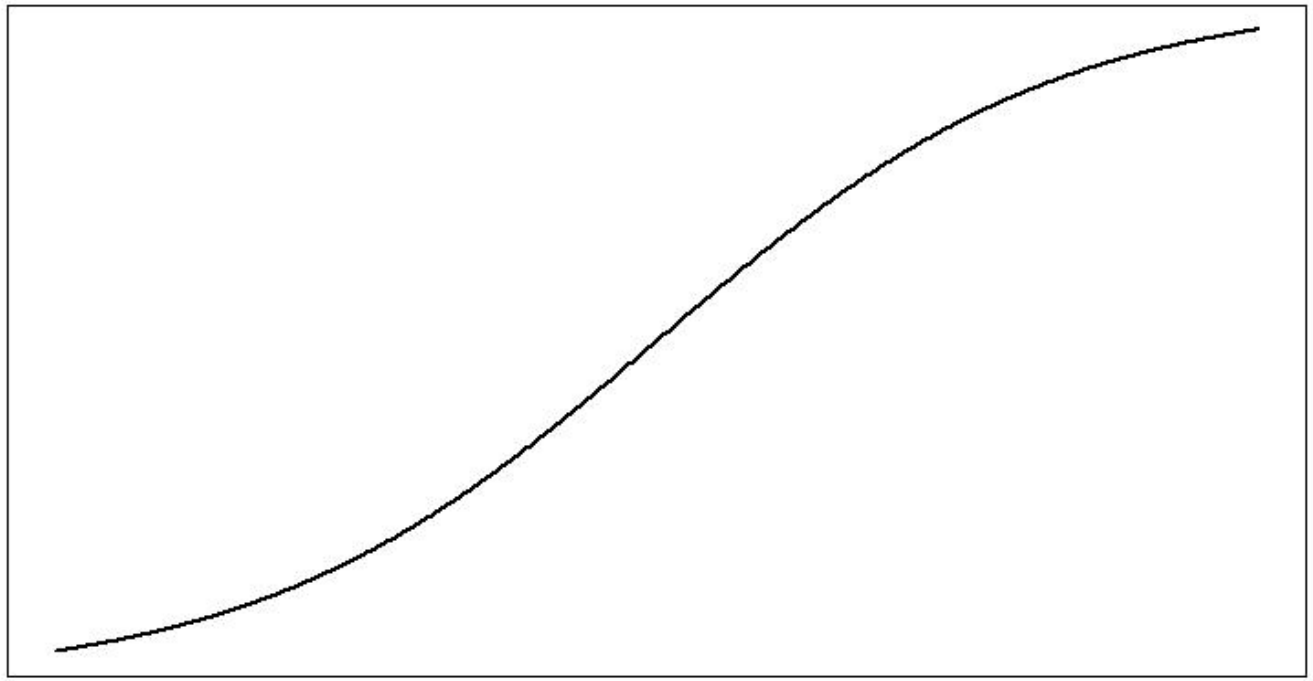
```
   survived  pclass     sex   age  sibsp  parch     fare embarked  class  ₩
0         0       3    male  22.0      1      0   7.2500        S  Third
1         1       1  female  38.0      1      0  71.2833        C  First
2         1       3  female  26.0      0      0   7.9250        S  Third
3         1       1  female  35.0      1      0  53.1000        S  First
4         0       3    male  35.0      0      0   8.0500        S  Third

     who  adult_male deck  embark_town alive  alone
0    man        True  NaN  Southampton    no  False
1  woman       False    C    Cherbourg   yes  False
2  woman       False  NaN  Southampton   yes   True
3  woman       False    C  Southampton   yes  False
4    man        True  NaN  Southampton    no   True
```

- 어떤 함수관계?
- $Y=f(X)$?
- $P(Y=1)=a+bX$?
- 해석은?
- 문제점은?

- X가 P(Y=1)에 미치는 영향이 선형이 아닐 수 있다.
- 예를 들어, 자동차 구매 종류 (신차,중고차) (Y)와 소득 (X)로 생각했을 때, 소득이 1000만원 증가 했을 때 신차 구매 확률에 미치는 영향은, (1)소득이 10억일 때 (P(Y=1)이 1에 가까울 것임)와 (2)소득이 3000만원 일 때 (P(Y=1)이 1보다 훨씬 낮음) 다를 것임.
- (1)일 때가 영향이 적음.
- 즉, S 모양의 관계가 있다고 볼 수 있음

- 어떤 S 모양의 함수를 쓸까?

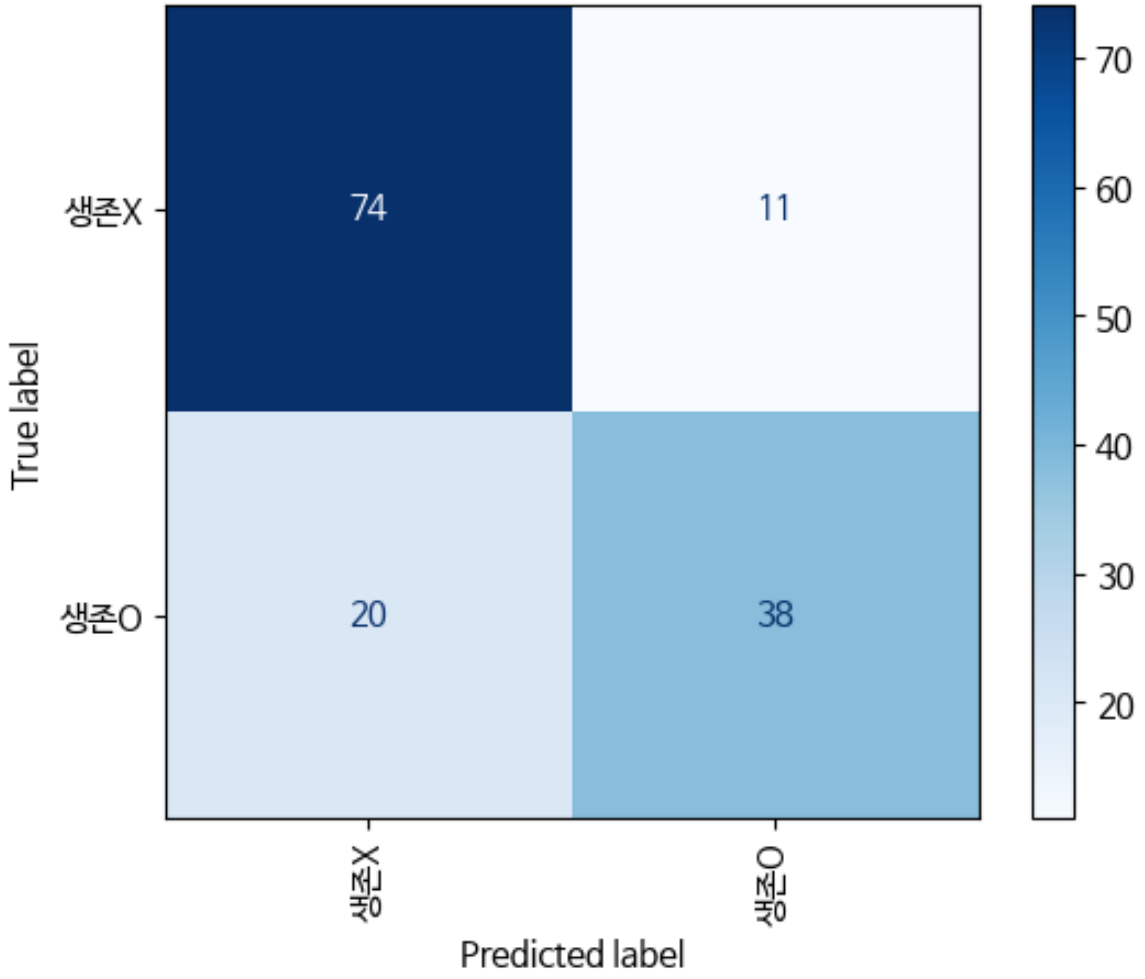


로짓함수

- logit(P(Y=1))=a+bX
- logit(x)= log(x/(1-x))
- P(Y=1)= exp(a+bX)/(1+exp(a+bX))
- X가 증가할 때, P(Y=1)은?
- $\hat{P}(Y=1) = \exp(\hat{a}+\hat{b}X)/(1+\exp(\hat{a}+\hat{b}X))$

- 모든 데이터들 (Y_i, X_i) 에 대하여 $\hat{P}(Y = 1|X = X_i)$ 를 계산하면 Y=1일 확률이 추정됨
- 이 값이 0.5이상이면 Y=1로 예측.
- 이 경우 $\hat{Y}_i = 1$로 놓자.
- 실제로 Y_i가 1인지 0인지 알기때문에 예측이 맞았는지 확인가능

- 생존여부(survived), 나이(age), 성별(sex), 캐빈등급(class)만 가지는 데이터 만들고, 결측치 제거
- 80%는 훈련용 데이터 (train), 20%는 검증용 데이터(test)로 분리
- 훈련용 데이터로 로지스틱회귀모형 적합



- 오분류율= (11+20)/143=0.22
- 민감도(Sensitivity)= 실제로 생존했을 때, 모형이 생존으로 예측하는 비율= 38/(20+38)=0.66
- 특이도(Specificity)= 실제로 생존하지 않았을 때, 모형이 생존이 아니라고 예측하는 비율= 74/(74+11)=0.87

- 생존여부 예측하기

- 3등석에 탑승한 30살 남성의 생존확률: 8.2%
- 2등석에 탑승한 20살 남성의 생존확률: 31.8%
- 1등석에 탑승한 10살 남성의 생존확률: 70.4%

마무리

데이터사이의 관계 찾기

- 함수관계를 가정하고, 데이터를 이용하여 통계적으로 추정할 수 있다.
- 추정된 함수 관계의 불확실성 (변동성)을 계산할 수 있다.
- 이를 통해, 함수관계가 통계적으로 유의미한지 판단할 수 있다.
- 연속적이 아닌 데이터의 경우도 함수관계를 데이터로 추정할 수 있다.

수고했어요!