Introduction to Statistics I

2025 Spring

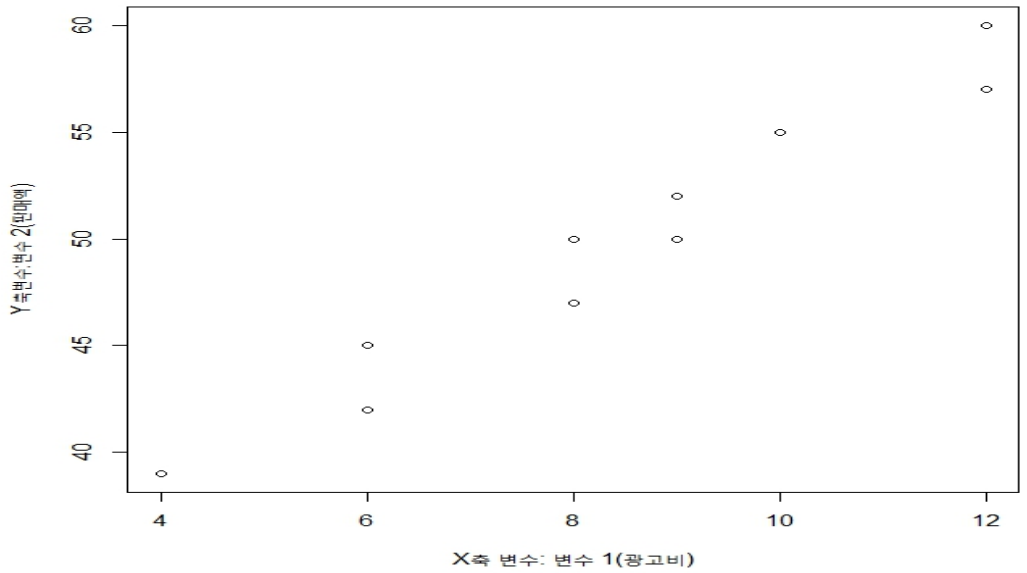
Industrial Data Science Lab & Unique AI

1. 상관 분석

- **상관분석**

회사	1	2	3	4	5	6	7	8	9	10
광고비(X)	4	6	6	8	8	9	9	10	12	12
판매액(Y)	39	42	45	47	50	50	52	55	57	60

- 광고비가 증가함에 따라 판매액도 증가
- 선형으로 증가

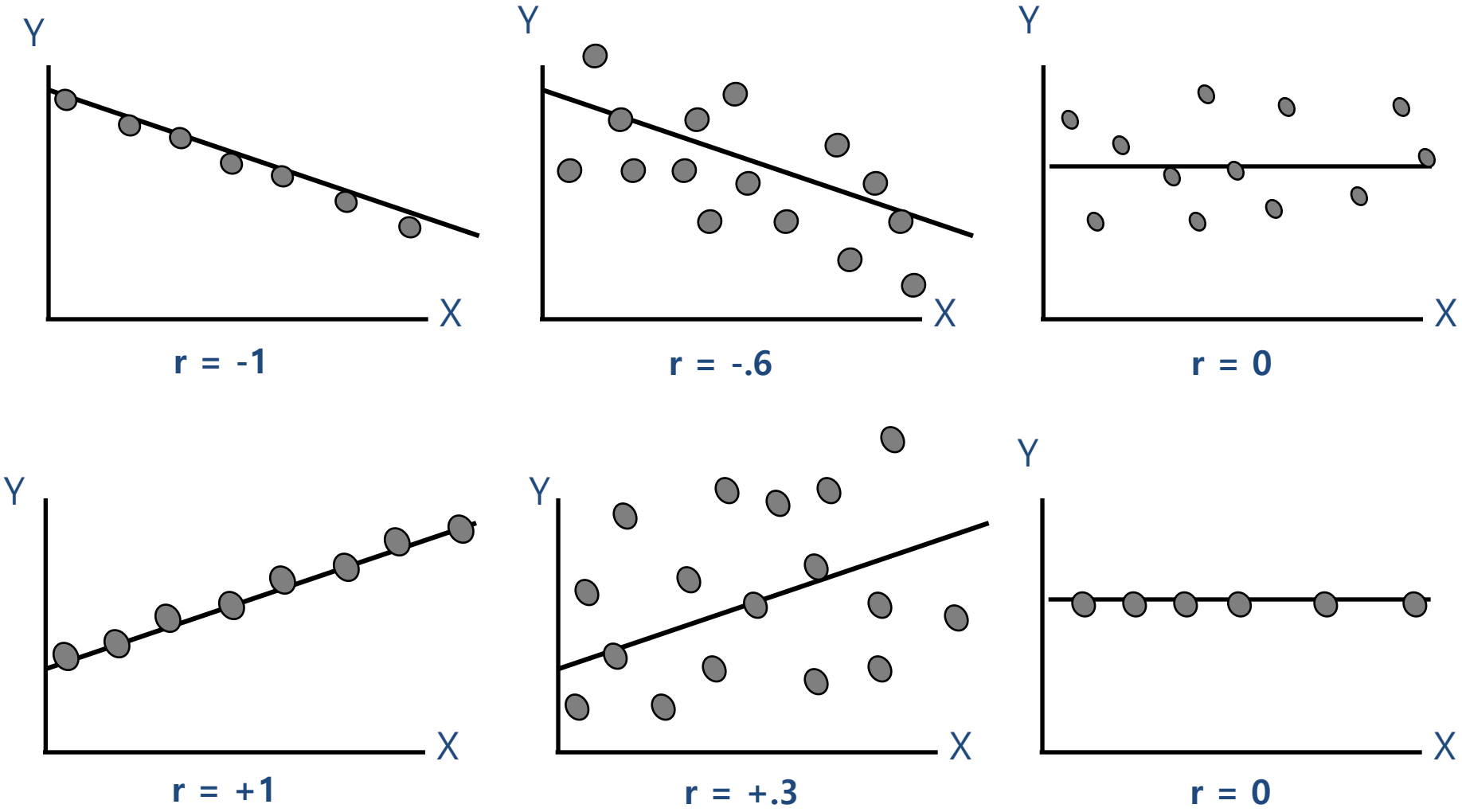


1. 상관 분석

- **공분산과 상관계수의 해석**
  - **공분산**
    - 두 변수의 상호 관계에 대한 강도를 나타냄
    - 두 변수의 원인/결과 관계를 나타내지 않는다.
      - Cov(x,y) > 0 x 와 y 는 같은 방향으로 움직이는 경향이 있음.
      - Cov(x,y) < 0 x 와 y 는 반대 방향으로 움직이는 경향이 있음
      - Cov(x,y) = 0 x 와 y 는 서로 독립이다.
  - **상관계수**
    - 두변수의 상호 관계를 나타내는 측도, 단위에 의존하지 않음
    - –1 과 1 사이의 값
      - –1 에 가까울수록 강한 음의 선형관계
      - 1 에 가까울수록 강한 양의 선형관계
      - 0 에 가까울수록 두 변수간의 선형관계는 없음 (다른 관계는 있을 수 있음)

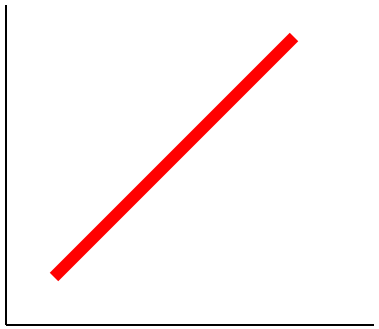
1. 상관 분석

- 산점도(Scatter plot)와 상관계수의 관계

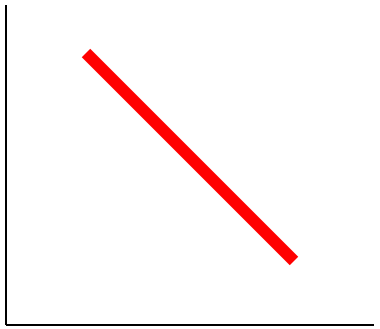


1. 상관 분석

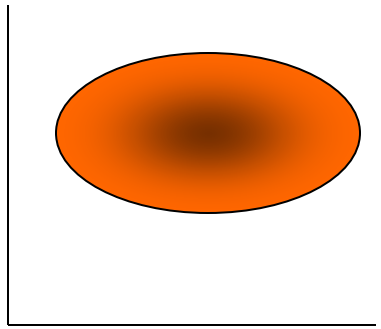
- **상관분석(Correlation Analysis)**



r = 1



r = -1



r = 0

- 상관관계 검정에 앞서 산점도를 이용하여 대략적인 관계를 파악
- 이상치 존재 여부 확인

1. 상관 분석

- **상관분석(Correlation Analysis)**
  - 두 변수(이변량 변수) 간 선형(상호의존)관계가 있을 경우 통계적으로 분석
    - 이변량 변수 : 같은 대상으로부터 측정된 두 변수
  - 상관계수 해석을 따름
    - 상관계수: 두 변수 간의 관련된 정도
    - 두 변수 간의 인과를 나타내지 않음

1. 상관 분석

- **상관분석(Correlation Analysis)의 예**

  - 목적 : 두 개의 집단 간에 직선적인 관계가 존재하는 지 데이터를 이용하여 파악
  - 사례 : 아버지와 딸의 키, 소득과 지출, 흡연량과 폐암, 공정 온도&강도
  - 가설
    - 귀무가설(H_0) : $\rho = 0$ (상관 관계가 없다.)
    - 대립가설(H_1) : $\rho \neq 0$ (상관 관계가 있다.)

  - 상관 계수(Correlation Coefficient ; r)
    - 두 집단간의 직선적인 관계를 나타내는 지표로서, -1부터 +1 사이 의 값을 갖음
    - -1에 가까울수록 "음의 상관"을 +1에 가까울수록 "양의 상관"을 갖음

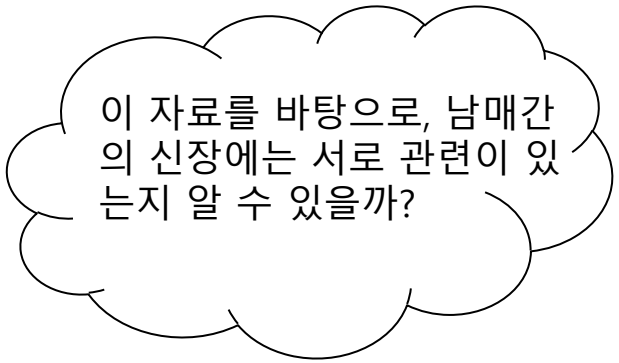
$$r = \frac{\sum(x_i - \bar{x})(y - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

1. 상관 분석

- **상관분석 예제(Correlation Analysis)**

남자키	여자키
71	69
68	64
66	65
67	63
70	65
71	62
70	65
73	64
72	66
65	59
66	62

성인 남매의 키를 측정한 자료



Pearson correlation of

남자 키 and 여자 키 = 0.558

P-Value = 0.074

2. 선형회귀분석 개요

- **Regression?**
  - 영국의 우생학자 Francis Galton (1822-1911)

  - 아버지와 아들의 키의 관계를 연구하며 Regression이라는 용어를 처음 사용
  - 아버지의 키가 큰 경우, 아들의 키는 작거나 아버지의 키가 작은 경우 아들의 키는 크며, 이들의 신장은 평균으로 가려는 경향
  - 부모의 키가 아들의 키에 영향을 주지만, 아들의 키는 그 세대 전체의 평균 신장으로 회귀

- **선형 회귀분석 (Linear Regression)**
- **목적 :**
  - 반응 인자(Response variable)와 하나 이상의 예측 인자(Predictor Variables) 사이의 관계를 표본으로부터 추정하여 수학적 모형을 만들고, 이를 통해 반응 인자에 대한 예측을 하는 방법

- **선형회귀는 데이터에 Straight line(기울기와 Y 절편)을 적합(fit)시키는 과정**
  - X변수들이 독립변수/Predictor 변수, Y변수가 종속변수/ Response 변수
  - 선형회귀를 통해 얻어진 Line은 기울기와 Y절편으로 나타내며, 알려진 X값에 대한 Y 값 예측

2. 선형회귀분석 개요

- 종류(인자 수에 의한 분류)
  - *단순회귀분석(Simple Linear Regression Analysis)*
    : 반응인자 1개와 예측 인자 1개로 구성 (예) $Y = b_0 + b_1X_1$
  - *다중 회귀분석(Multiple Regression Analysis)*
    : 반응인자 1개와 두 개 이상의 예측 인자로 구성 (예) $Y= b_0 + b_1X_1 + b_2X_2 + \dots + b_kX_k$

- Residual(잔차): 이렇게 예측된 Y와 실제 Y의 차이
  - 관측치가 Straight line에서 얼마나 떨어져있는지를 나타냄
  - Least squares : 기울기와 Y절편인 b0 와 b1 를 구하는 방법, Residual errors의 Square의 합을 최소화하는 기울기와 절편을 찾음

- 유의사항
  - 통계적 추론을 하기 위해서는, 잔차에 대한 가정이 필요
  - 등분산성, 선형성, 정규성 (선형성은 변수와 잔차의 Scatter plot 이용하여 확인)

2. 선형회귀분석 개요

- **선형회귀 검정**
  - 회귀 계수에 대한 검정
  - 회귀계수=0이면 그 회귀계수에 해당하는 독립변수는 종속변수와 관계가 없다고 해석
  - 회귀계수=0을 H0으로 보고, 다음의 통계량을 통해 검정

$$t = \frac{\hat{b}}{s.e.(\hat{b})}$$

  - 검정통계량에 대한 확률을 구해서 그 확률이 유의수준보다 작은 경우 H0을 기각
  - 회귀계수는 0이 아니며, 구해진 회귀계수는 유의함

2. 선형회귀분석 개요

- **선형회귀분석**
  - 선형회귀분석의 변수
    - 독립변수(Independent variable : X)
      - 종속변수에 영향을 주는 변수(설명변수. 외생변수 등)
      - 종속변수를 설명하는데 이용하는 변수
      - 예 : 광고비와 판매액의 관계에서 광고비
    - 종속변수(Dependent variable : Y)
      - 서로 관계를 가지고 있는 변수들 중에서 다른 변수에 의해 영향을 받은 변수(반응변수, 내생변수 등)
      - 설명하고자 하는 변수
      - 주로 다른 변수의 반응으로 관측되는 변수
      - 예 : 광고비와 판매액의 관계에서 판매액
  - 회귀분석(Regression analysis):하나 이상의 독립변수를 이용하여 종속변수를 예측하는 통계적 분석
    - 단순선형회귀분석 : 독립변수가 하나인 경우
    - 다중선형회귀분석 : 독립변수가 2개 이상의 경우

2. 선형회귀분석 개요

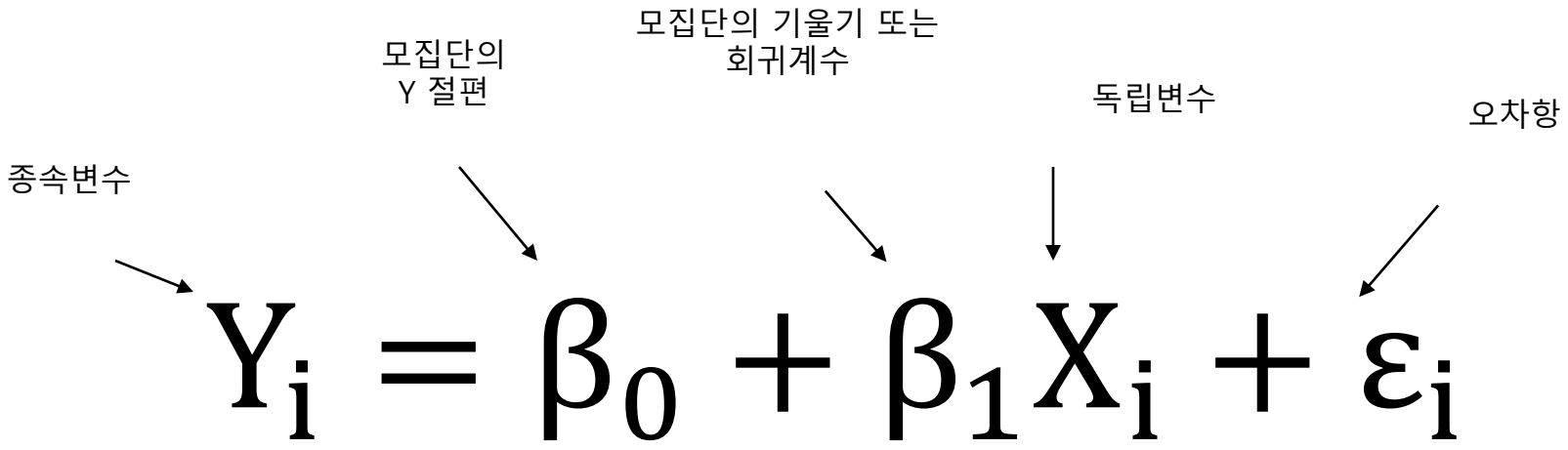
- **단순선형회귀모형**
  - X 와 Y의 관계를 선형함수(선형 방정식)로 표현
  - 단순선형회귀 모형

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

  - β_0 와 β_1: 미지의 모수, 회귀계수(Regression coefficients)
  - ε : 회귀식으로 표현할 수 없는 오차항

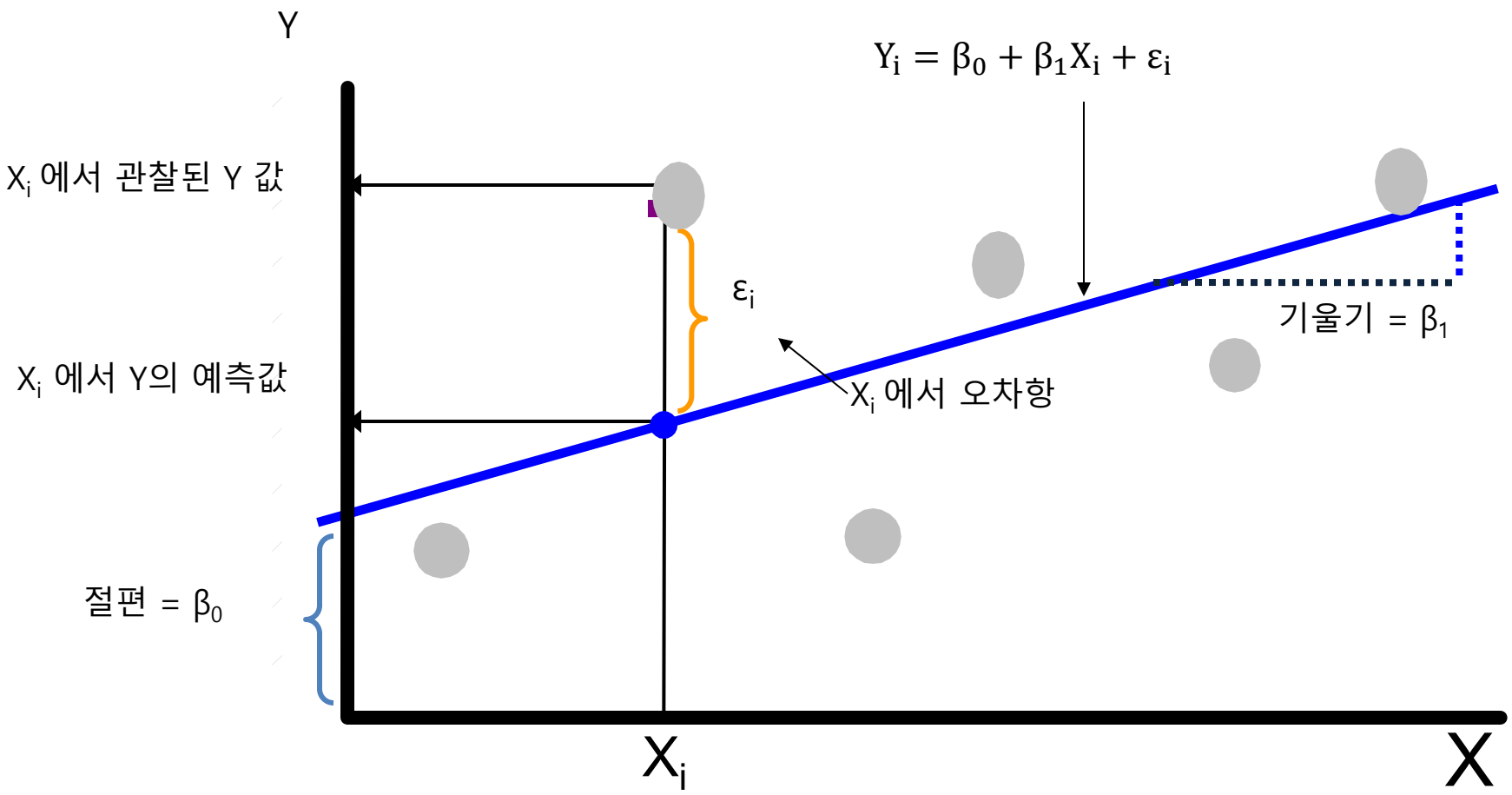
3. 선형회귀분석

- **단순선형회귀모형**



3. 선형회귀분석

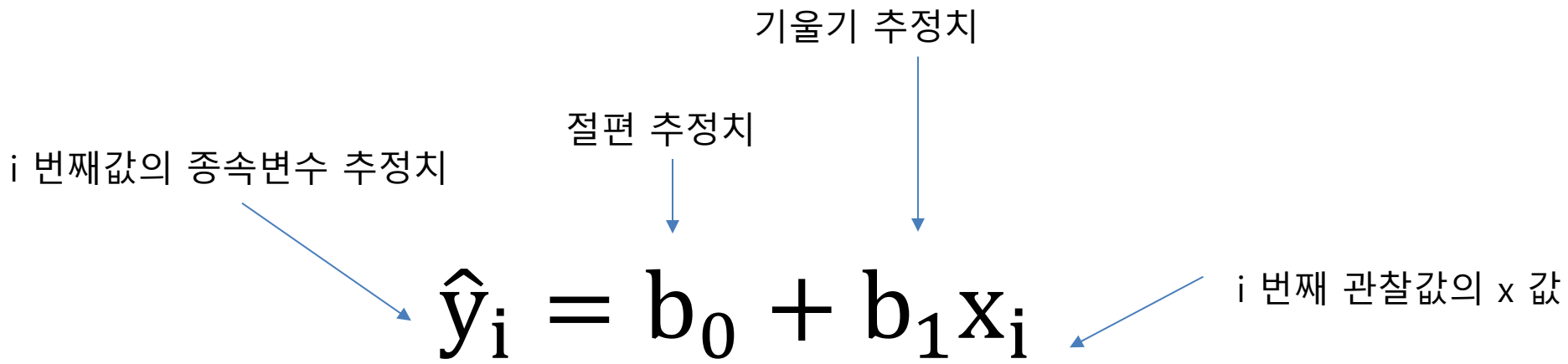
- **단순선형회귀모형**



3. 선형회귀분석

- **회귀 모형 기본 가정**
  - X 와 Y는 오차항(ε)을 포함하는 선형관계(Linear Relation)
    - 오차항 (ε_i) 의 평균은 0
      - $E(\varepsilon_i)=0$ (i=1, 2, ….,n)
    - (등분산성) 모든 ε_i 의 분산은 σ^2 으로 동일
      - $Var(\varepsilon_i)= \sigma^2$ (i=1, 2, ….,n)
    - (독립성) 오차항 ε_i 들은 서로 독립
      - 서로 다른 i, j 에 대하여 $E(\varepsilon_i\ \varepsilon_j)=0$
    - (정규성) 오차항 ε_i 들은 정규 분포
      - ε_i ~N(0, σ^2) (i=1, 2, ….,n)

- **회귀계수 추정**
  - 표본회귀식 : 표본을 이용하여 추정된(적합된) 회귀식



  - 잔차 e (residual) : 예측된 값과 실제 관찰값의 차이

$$e_i = (y_i - \hat{y}_i) = y_i - (b_0 + b_1 x_i)$$

- **최소제곱법 (Least Square Estimation)**
  - 잔차: 적합된 회귀식에서 계산된 예측치($\hat{y}_i$)와 관찰치(y_i)의 차이
  - LSE: 잔차들의 제곱의 합이 최소가 되도록 회귀계수를 추정하는 방법
    - 다음의 식을 최소화 하는 b_0 와 b_1의 값을 구하기

$$\min \ \mathrm{SSE} = \min \sum e_i^2$$

$$= \min \sum (y_i - \hat{y}_i)^2$$

$$= \min \sum [y_i - \ (b_0 + b_1 x_i)]^2$$

위의 식을 b_0 과 b_1 에 대하여 편미분하여 0으로 놓고
b_0 과 b_1 에 대하여 풀어서 구해지는 값

- **최소제곱 추정량**
  - 기울기에 대한 추정량

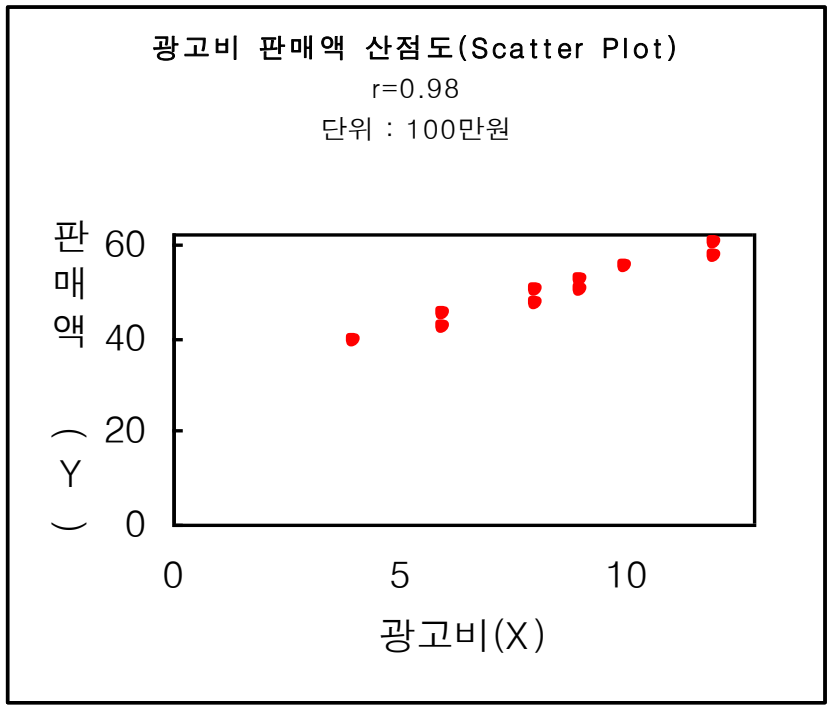
$$b_1 = \frac{s_{XY}}{s_X^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = r\ \frac{s_Y}{s_X}$$

  - 절편에 대한 추정량

$$b_0 = \bar{y} - b_1\bar{x}$$

- **단순선형회귀모형: 기울기와 절편에의 해석**
  - b0 : x 가 0 의 값을 가질 때 y 값의 추정량
  - b1 : x 가 한 단위 증가할 때 y 의 증가분

판매액 (Y)	광고비 (X)
39	4
42	6
45	6
47	8
50	8
50	9
52	9
55	10
57	12
60	12



3. 선형회귀분석

- **단순선형회귀모형**

  - 기울기에 대한 추정값

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{151.2}{60.4} = 2.503$$

  - 절편에 대한 추정값

$$b_0 = \bar{y} - b_1\bar{x} = 49.7 - 2.5033 \times 8.4 = 28.627$$

  - 적합된 회귀식

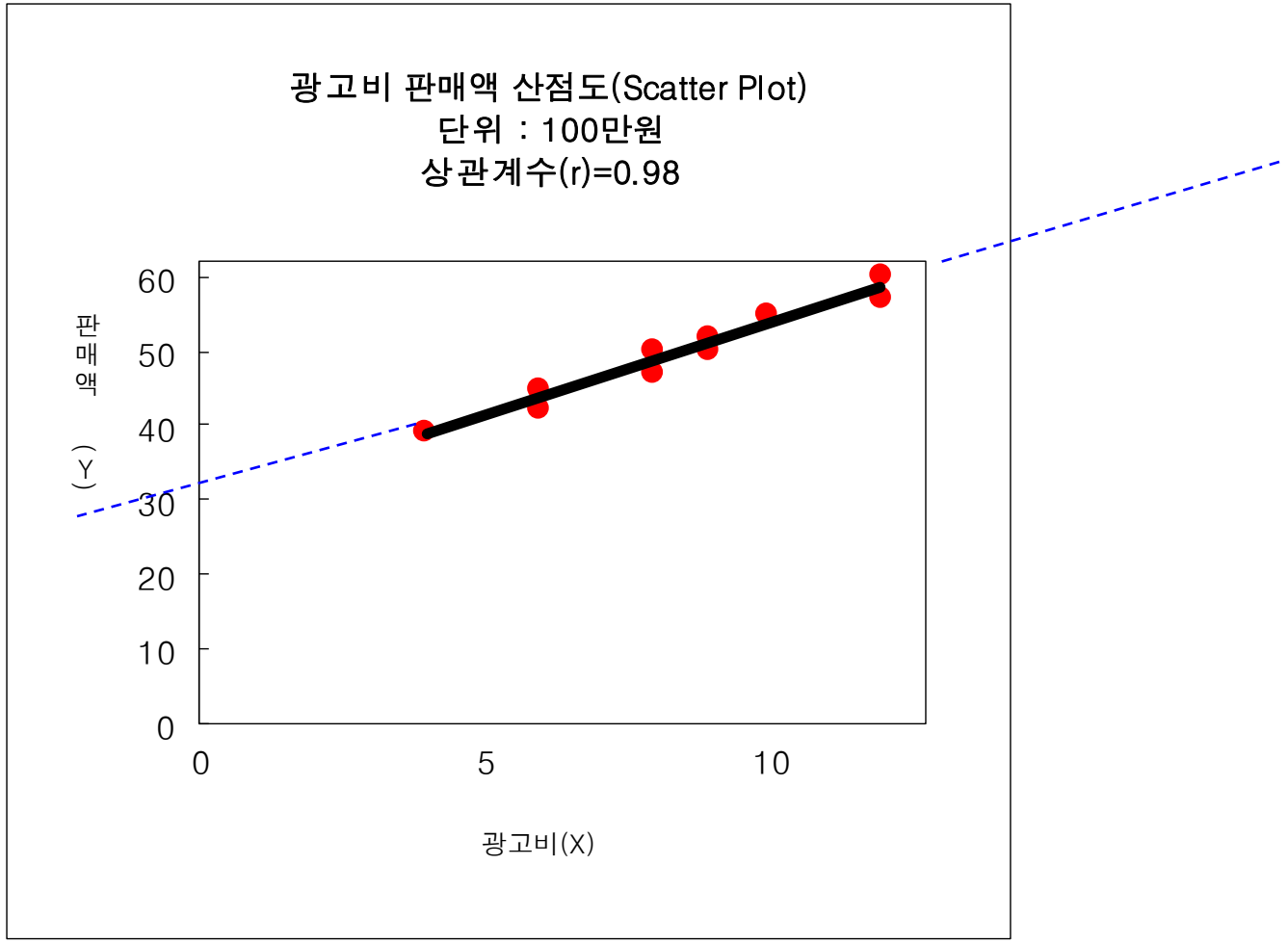
$$\hat{y}_i = b_0 + b_1 x_i = 28.627 + 2.5033 x_i$$

- b0 =28.627 :
- 광고비가 0일때 판매액은 약 2천8백만원 예상

- b1=2.5033 :
- 광고비가 1단위 증가 시 판매액 평균적으로 약 2.5백만 원 증가

3. 선형회귀분석

- **단순선형회귀모형**



- **회귀직선의 적합도**
  - Y의 관측값들이 가지는 총변동 SST

$$SST = SSR + SSE$$

총 제곱합

$$SST = \sum (y_i - \bar{y})^2$$

회귀제곱합

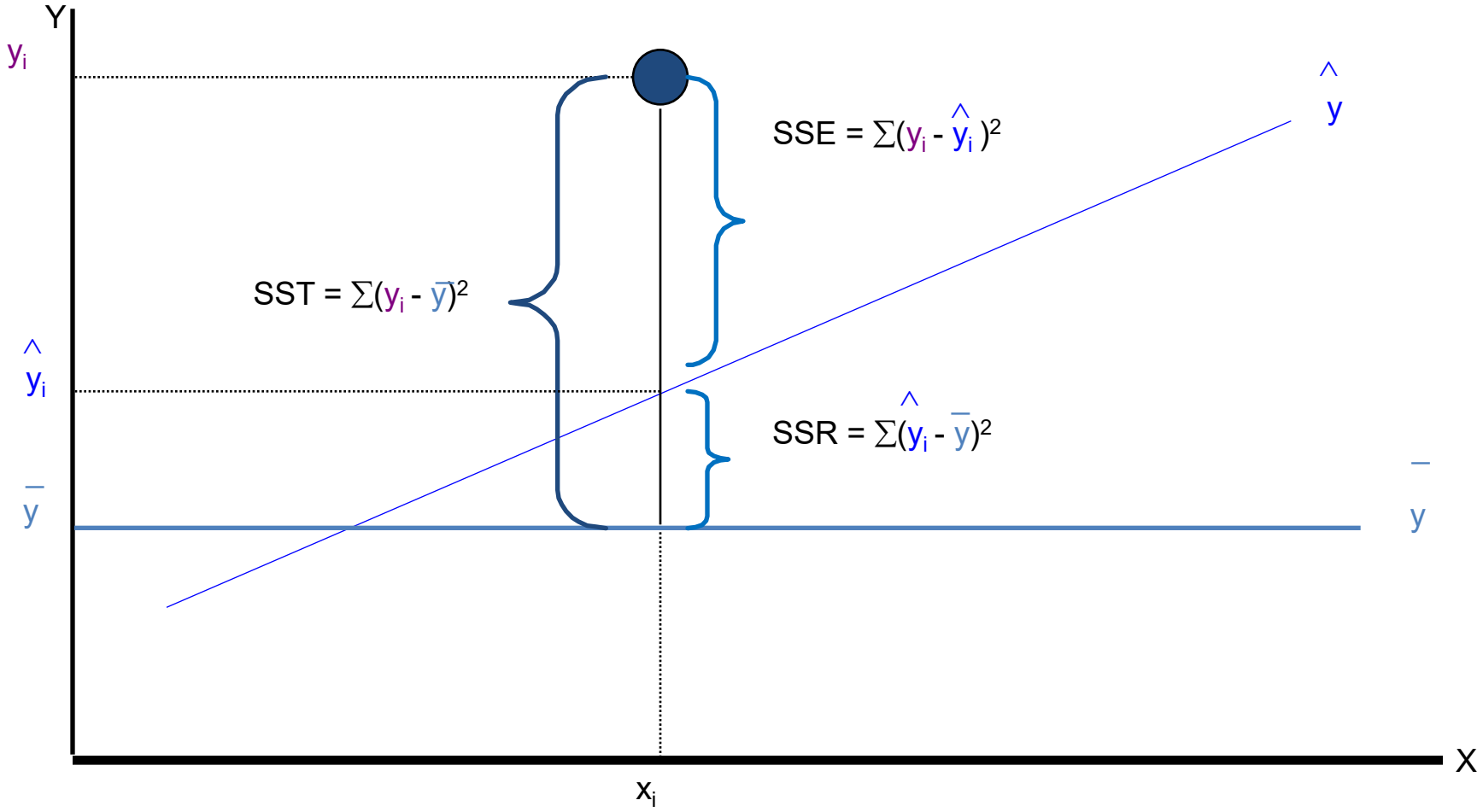
$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

오차제곱합

$$SSE = \sum (y_i - \hat{y}_i)^2$$

$\bar{y}$: 종속변수의 평균
y_i : 종속변수의 관찰값
$\hat{y}_i$: x_i 에 대한 종속변수의 추정값

3. 선형회귀분석



- SST = total sum of squares , Y의 관측값들이 가지는 총변동을 나타내는 제곱합
- SSR = regression sum of squares, Y의 총변동 중 회귀식에 의해 설명되는 변동을 나타내는 제곱합
- SSE = error sum of squares , 잔차들의 제곱합으로 Y의 총변동 중 회귀식에 의해서 설명되지 않는 변동

- **결정계수(Coefficient of Determination)**

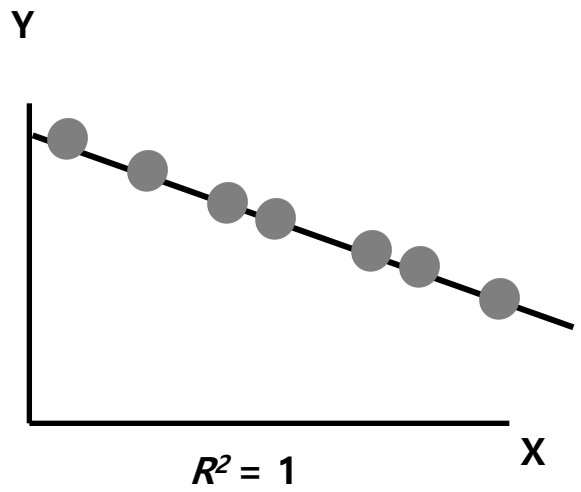
  - 결정계수(R^2) : 총변동 SST 중에서 회귀에 의해 설명된 변동 SSR이 차지하는 비

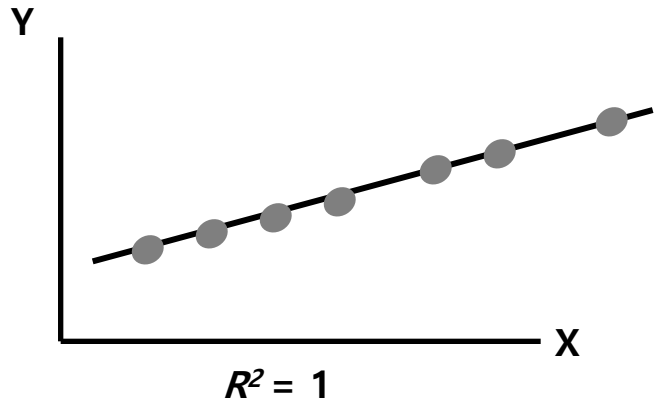
$$R^2 = \frac{SSR}{SST} = \frac{\text{회귀식에 의해 설명된 변동}}{\text{총 변동}}$$

  - 결정계수의 값은 항상 0과 1 사이
    - 1에 가까울 수록 표본들이 회귀직선 주위에 밀집되어 있음
    - 추정된 회귀식이 관측값을 잘 설명하고 있음

$$0 \leq R^2 \leq 1$$

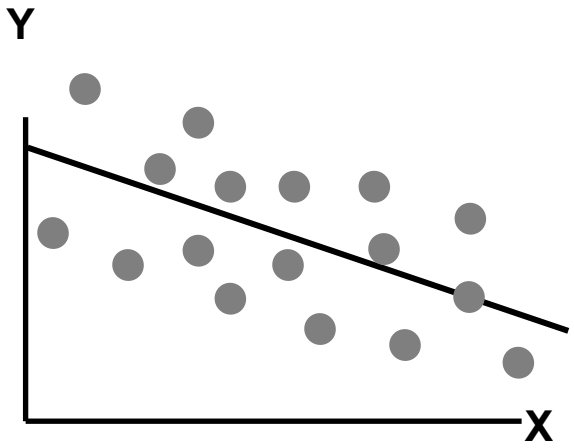
➢ 결정계수

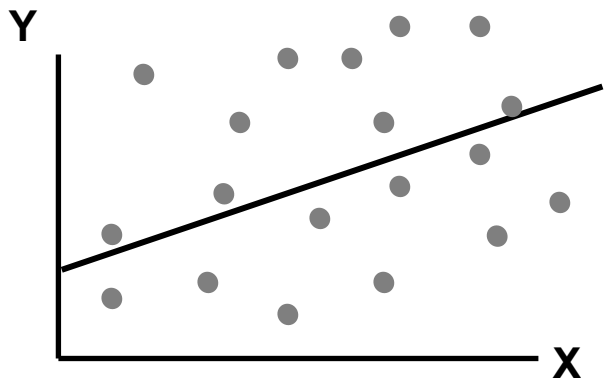




- 선형관계
- Y는 X에 의해서 100% 설명

- **결정계수**

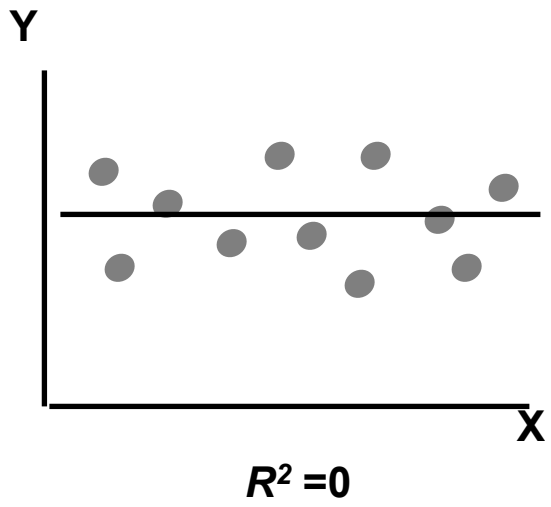




$0 < R^2 < 1$

- X 와 Y 의 선형관계
- Y는 X에 의해 r^2 X100% 설명

- 결정계수



$R^2 = 0$

- X 와 Y의 선형관계가 없음
- Y는 X에 의해 설명되지 않음

$R^2 = r_{xy}^2$

- 결정계수는 X와 Y의 표본상관계수의 제곱과 일치

3. 선형회귀분석

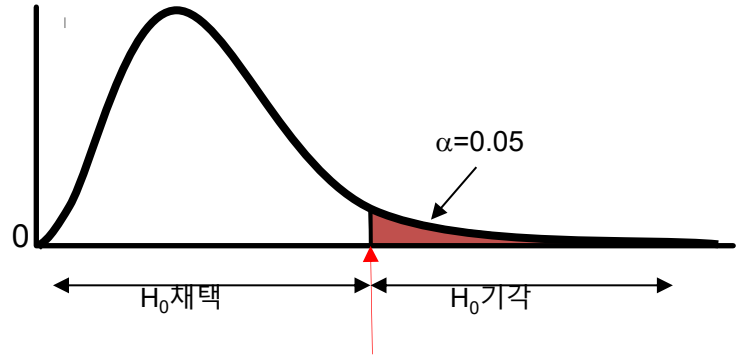
- **회귀분석의 분산 분석표**

요인	제곱합(SS)	자유도(df)	평균제곱(MS)	F비
R(회귀)	SSR	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
E(오차)	SSE	n-2	$MSE = \frac{SSE}{n-2}$	
T(전체)	SST = SSR+SSE	n-1		

- **가설검정**

$$F = \frac{MSR}{MSE} \sim F(1, n-2)$$

$H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$



- 귀무 가설 기각
  - 만약 β_1 이 0 이 아니라면 가정된 회귀식이 타당하여 Y의 변동이 회귀식에 의해 설명, 이 경우 F값이 크게 됨
  - F 값이 충분히 크면 β_1 이 0 이 아니라고 해석

- 귀무가설이 맞는 경우
  - 검정통계량 F는 자유도가 (1,n-2)인 F 분포
  - $F > F_{1, n-2, \alpha}$이면 귀무가설 기각

- **기울기 (β1)에 대한 가설 검정**
  - 모집단의 기울기가 0 인지 ?
    - H_0: $\beta_1 = 0$(선형관계가 없음)
    - H_1: $\beta_1 \neq 0$(선형관계가 존재)

$$|t| = \left|\frac{b_1-\beta_1}{SE(b_1)}\right|$$ 와 유의수준a에서의 통계량값과 비교$(t_{n-2,\alpha/2})$

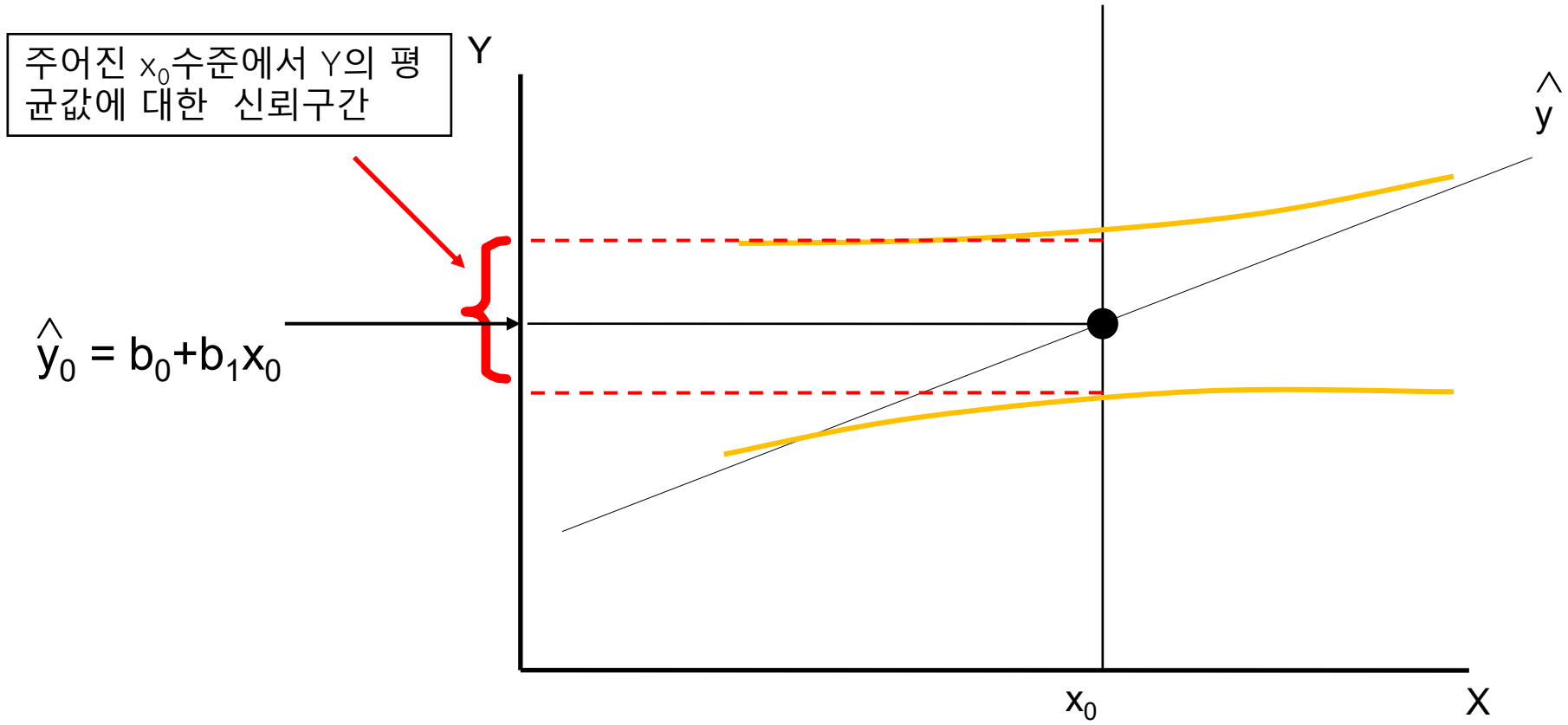
  - 모집단의 절편이 0 인지
    - H_0: $\beta_0 = 0$
    - H_1: $\beta_0 \neq 0$

$$|t| = \left|\frac{b_0-\beta_0}{SE(b_0)}\right|$$ 와 유의수준a에서의 통계량값과 비교$(t_{n-2,\alpha/2})$

3. 선형회귀분석

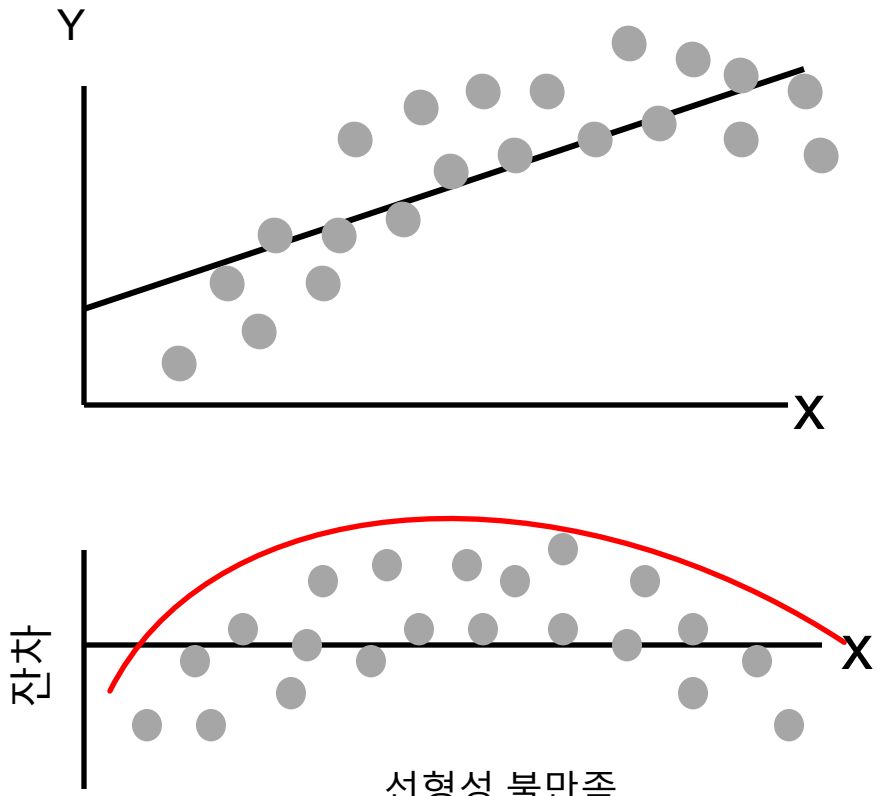
- **Y의 평균값 E(Y|x_0)=$\mu_{Y|x_0}$**
  - X의 임의의 점 X=x0 에서, 종속변수 Y는 평균값 E(Y| x_0)= $\beta_0 + \beta_1 x_0$
  - X=x_0인 동일 수준에서 Y의 평균값을 의미
  - E(Y| x_0)에 대한 점 추정량
    - $\hat{y}_0 = b_0 + b_1 x_0$
    - $\hat{y}_0$의 신뢰 구간 : $\hat{y}_0 \pm t_{n-2,\frac{\alpha}{2}} SE(\hat{y}_0)$
    - 주어진 X의 SE가 크면 신뢰구간의 폭도 넓어짐
    - 신뢰대(Confidence Band): X의 각 점에서 Y평균에 대한 신뢰 구간을 구하여 상/하한을 연결한 곡선

- **Confidence Band**



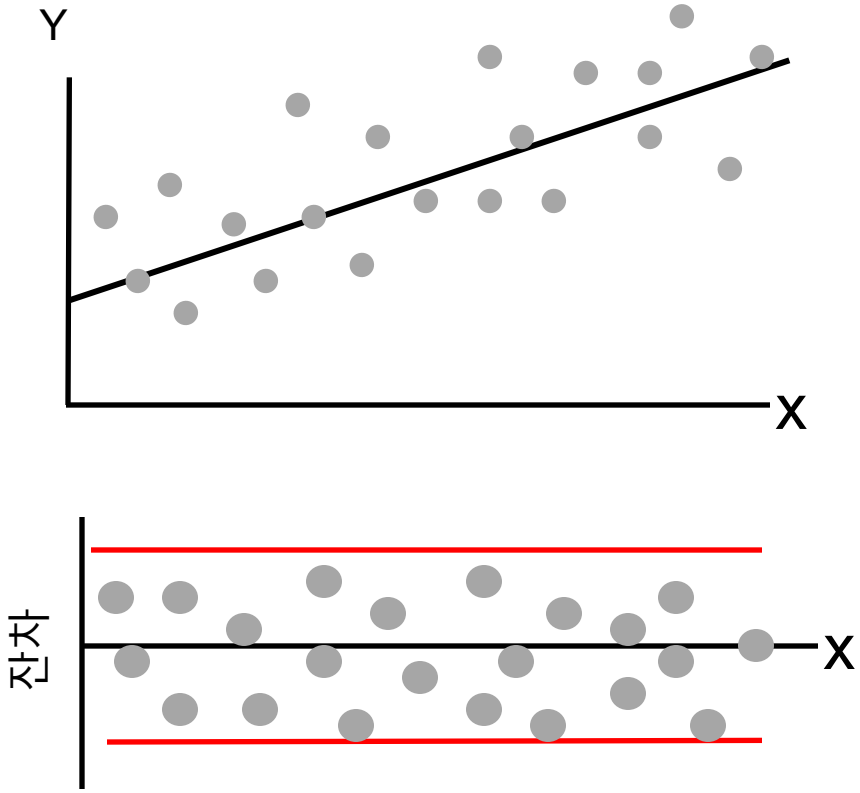
- **잔차 분석(Residual Analysis)**
  - 데이터가 회귀 분석의 기본 가정들을 만족하는지 잔차 e의 산점도로 분석
    - 자료의 선형성: '잔차 vs 독립변수' 산점도
    - 오차항의 등분산성: '잔차 vs 예측치' 산점도
    - 오차항의 독립성과 $E(\varepsilon_i)=0$: '잔차 vs 관측순서' 산점도 주로 이용.
    - 오차항의 정규성 : 잔차의 히스토그램 확인, Q-Q plot 등
  - 위의 산점도에서 잔차들이 0을 중심으로 랜덤하게 나타나면 각 가정을 만족

- 잔차 분석(Residual Analysis) - 선형성 점검



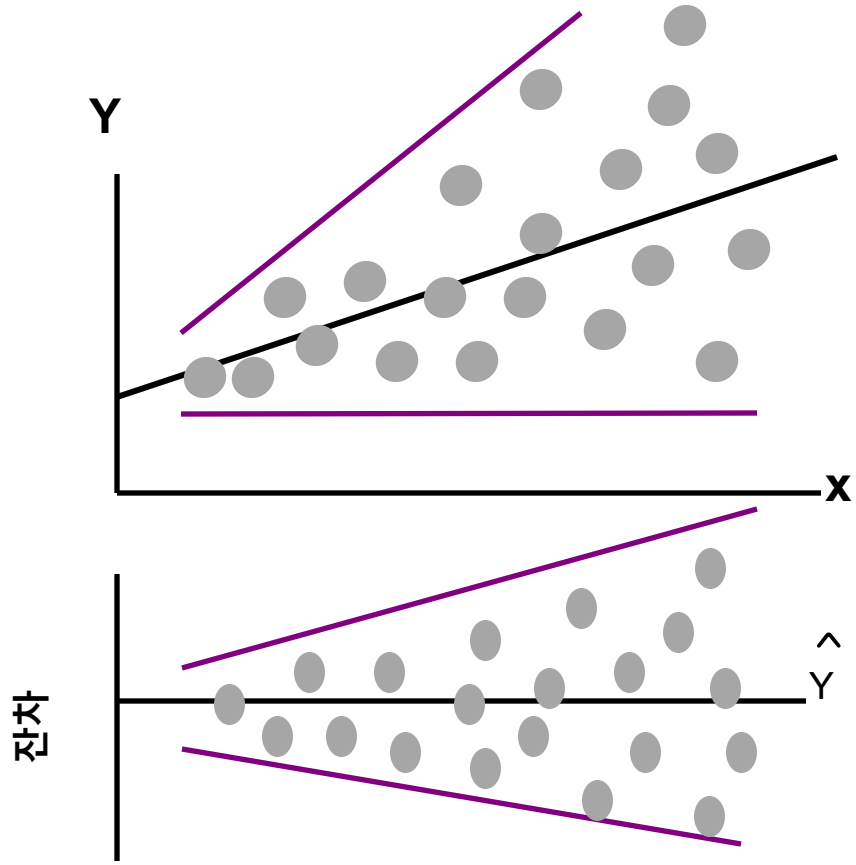
선형성 불만족

(비선형)

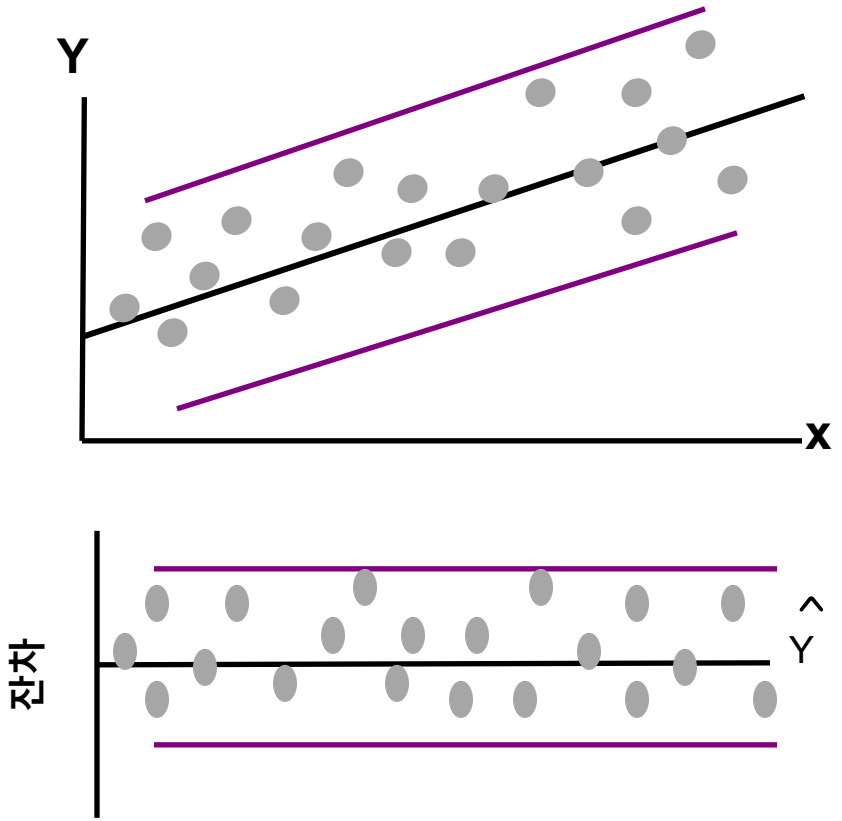


선형성 만족

- 잔차 분석(Residual Analysis) - 등분산성 점검

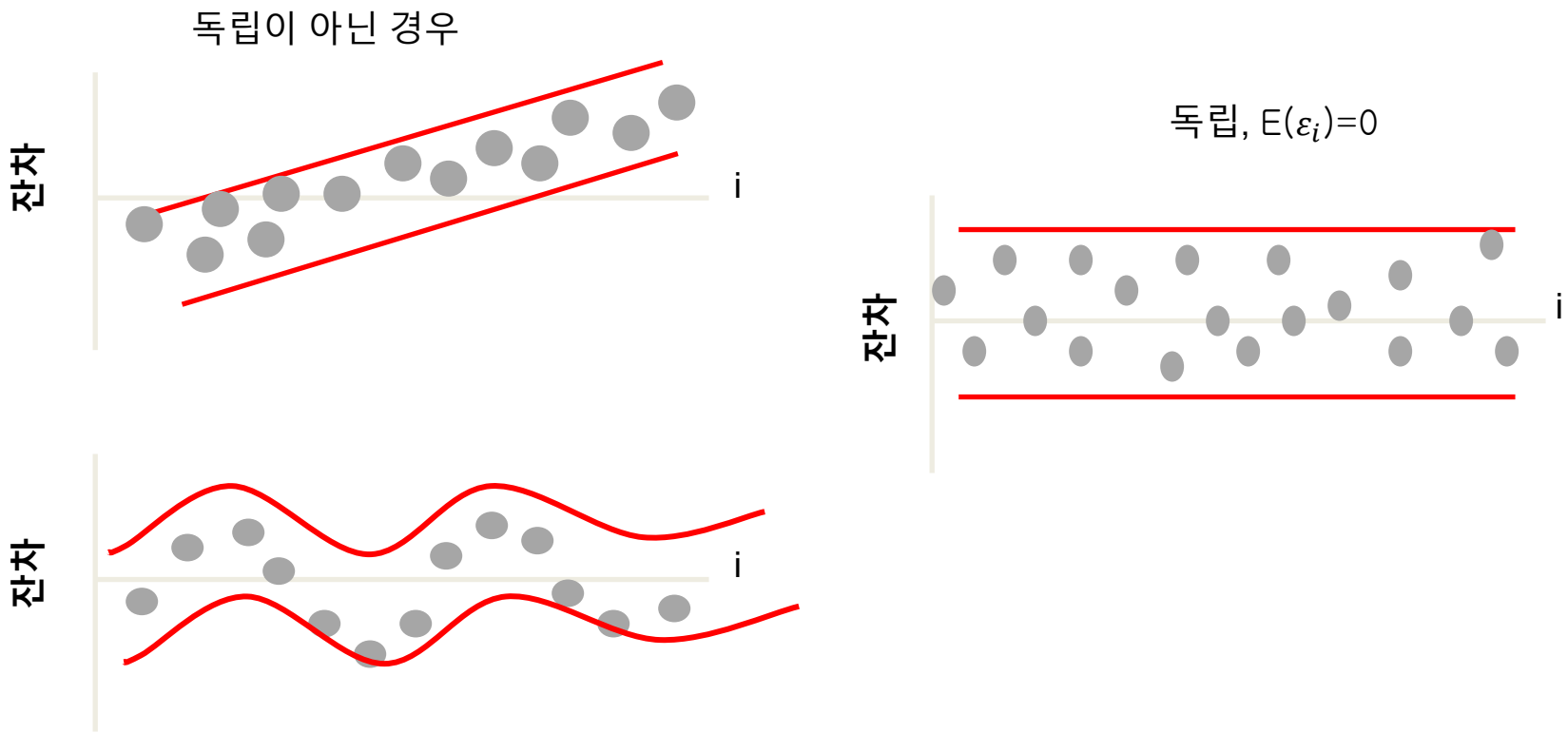


이분산



등분산

- **잔차 분석(Residual Analysis) - 오차항의 평균=0, 오차항은 서로 독립에 대한 점검**



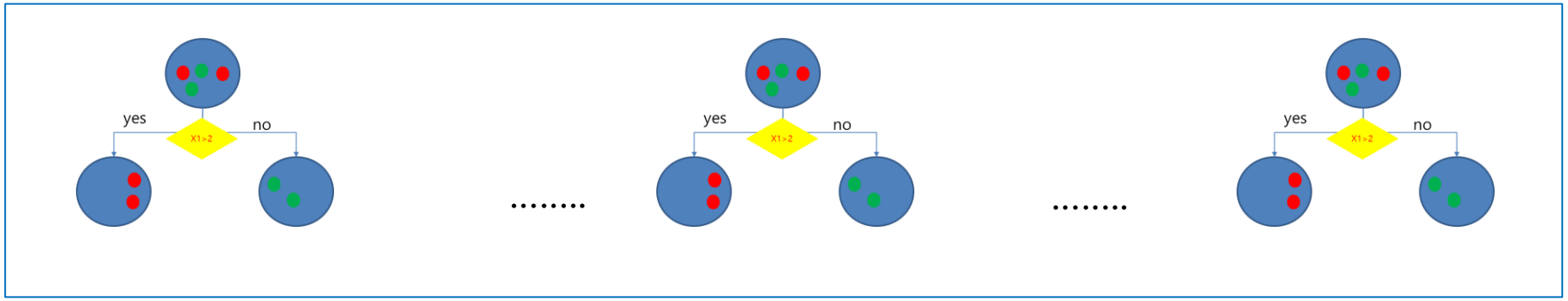
4. Ensemble

- Ensemble 기법: 여러 분류 모형의 결과를 결합하는 기법
- Random Forest: 앙상블 학습 방법의 일종으로, 훈련 과정에서 구성한 다수의 Decision Tree로부터 Voting을 통해 결과 예측
- Bagging: 주어진 데이터에서 랜덤하게 여러 개의 같은 크기의 부분집합을 생성
- Out of Bag과 Voting: Out of Bag(OOB)는 Bagging에서 제외되는 데이터들을 의미하며, Voting은 Random Forest내 여러 Decision Tree의 결과 중 다수의 결과를 선택하는 방법

4. Ensemble

Random Forest

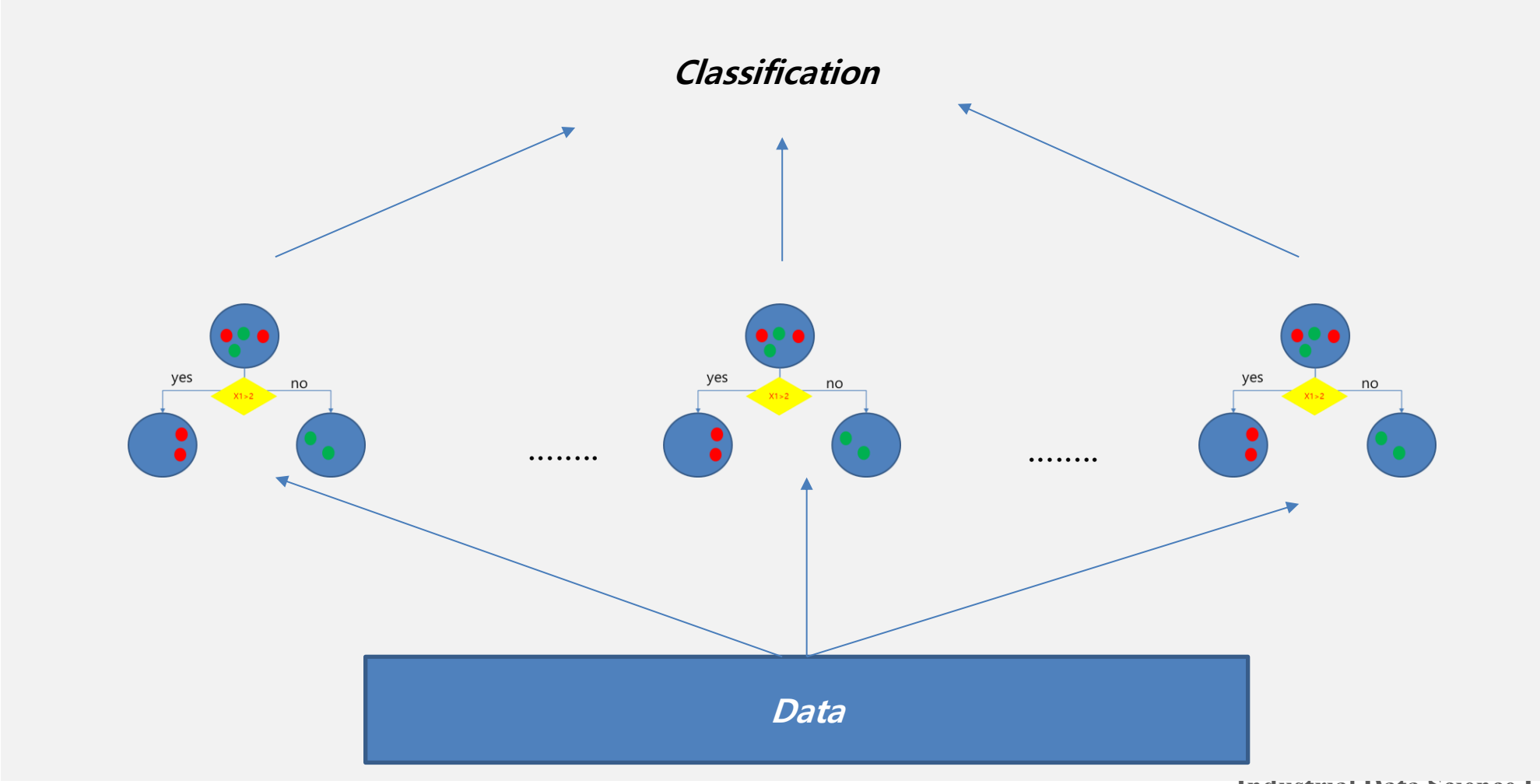
- Breiman의 " bagging " 과 변수 랜덤 선택 아이디어 기반
- 처음에는 random decision forests로 시작하여 발전
- 데이터의 다양한 경우를 반영할 수 있도록 보완
- 다양한 경우에 대한 Decision Tree를 통해 성능과 안정성을 제고



4. Ensemble

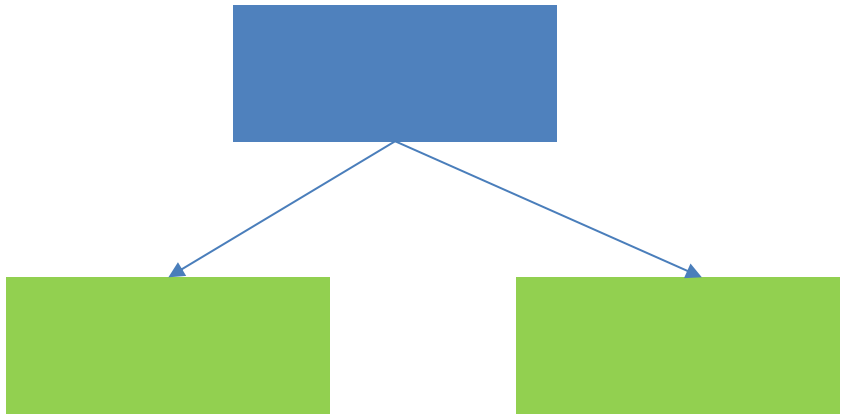
Random Forest

- 몇 개의 Decision Tree를 만들 것인지?
- 몇 개의 X변수를 Random하게 선택할 것인지?



4. Ensemble

- Adaboost는 Ensemble 기법의 Boosting을 DT에 적용
- Stump로 부터 학습을 시작
  - Stump: 단순한 형태의 Tree, Weak learner

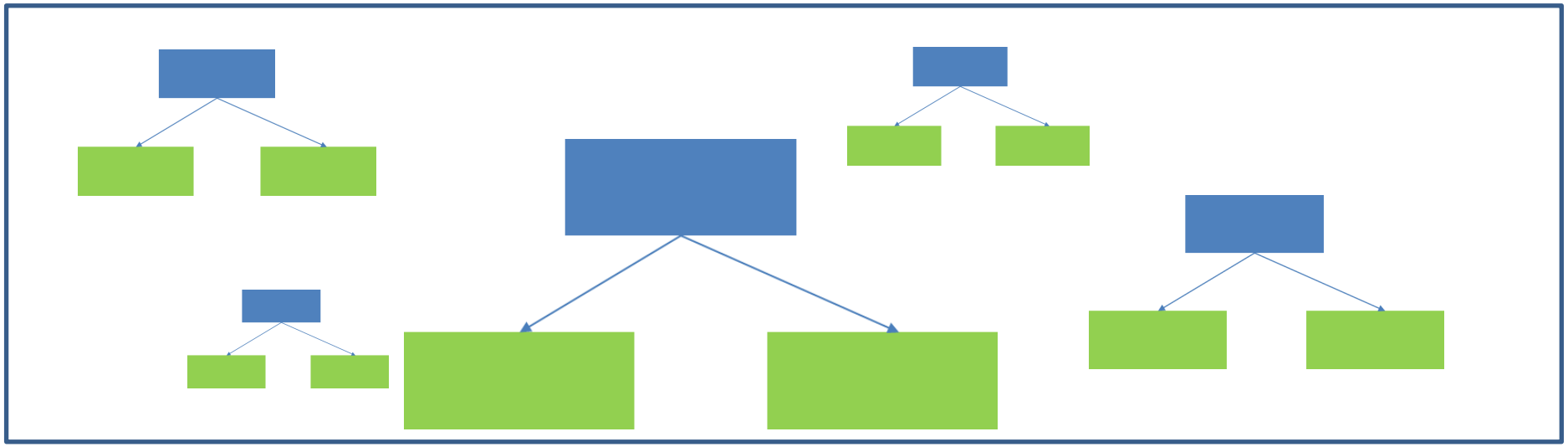


4. Ensemble

- Forest of stumps를 활용
  - Random Forest: 모든 tree는 같은 weight을 갖음
  - Adaboost: Stump마다 중요도의 차이가 존재

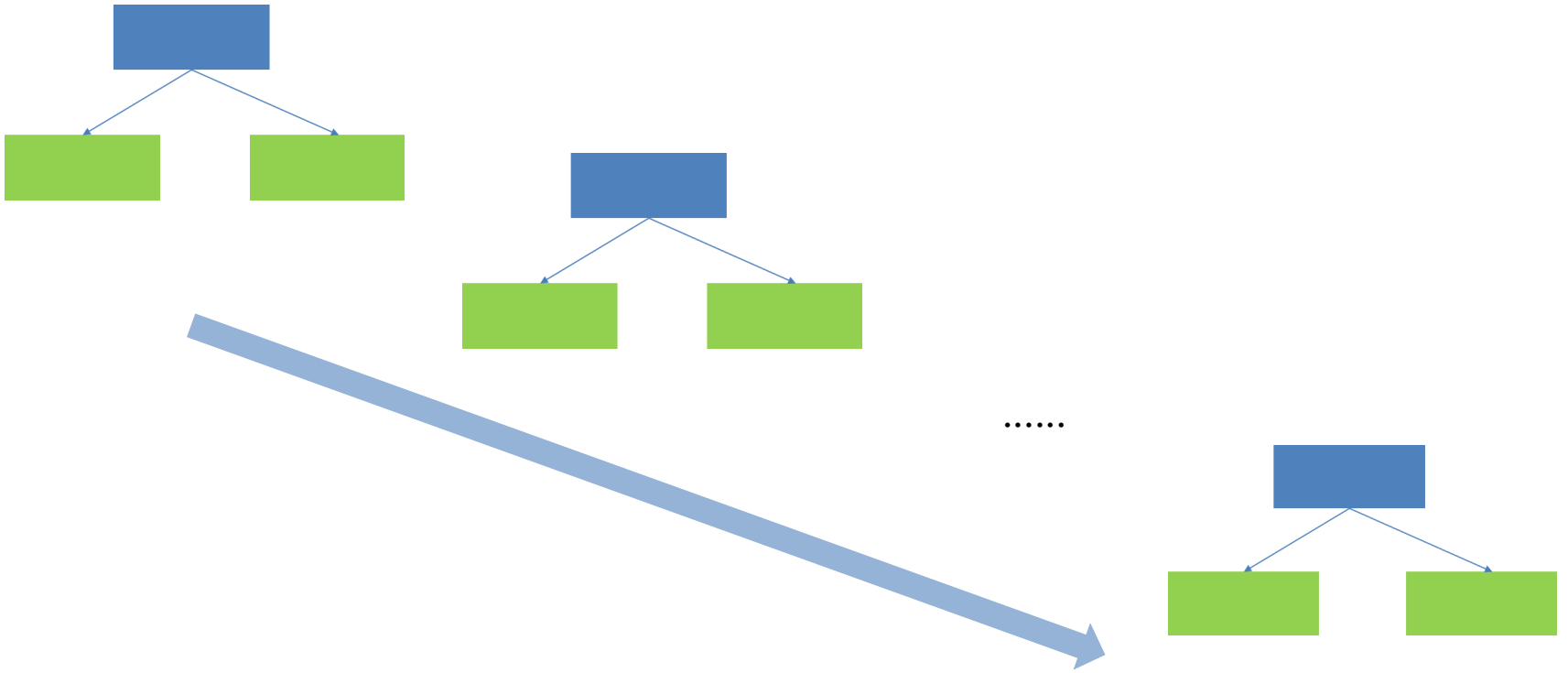
- Random Forest에서는 Tree가 같은 중요도를 지님

- Adaboost에서의 stump의 중요도: Amount of say로 표현, 클 수록 결과에 큰 영향을 미침



4. Ensemble

- Forest of stumps
  - 첫 stump는 다음 stump에 영향, 순차적으로 다음 stump에 영향을 주는 방식



- **Gradient Boost for Regression**
  - GB: leaf로 부터 시작
    - Leaf: Target에 대한 초기 추정값(예: 평균, log(odds ratio) 등)
    - Stump가 아닌 Tree를 생성: 각 tree는 leaf가 8~32개 크기 수준으로 생성

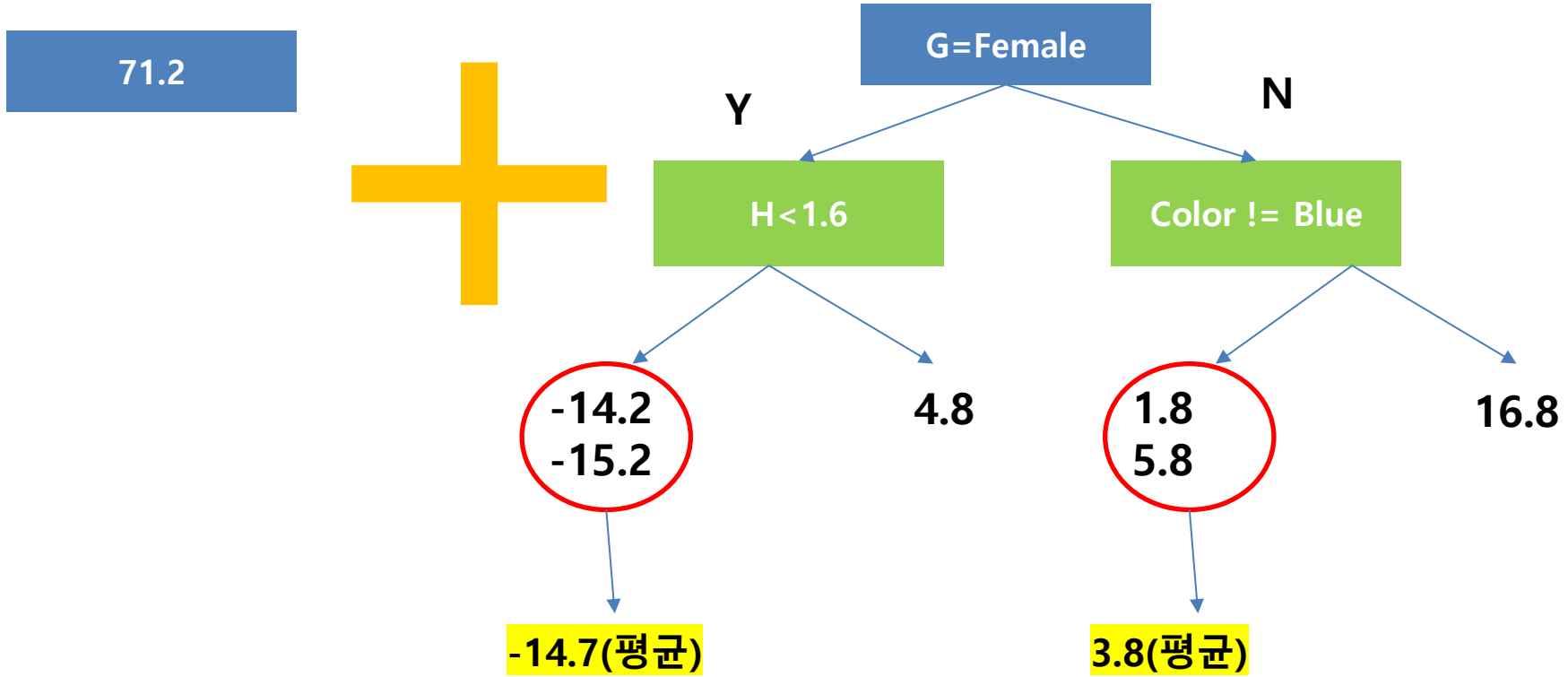
			Target
Height	Color	Gender	Weight
1.6	B	M	88
1.6	G	F	76
1.5	B	F	56
1.8	R	M	73
1.5	G	M	77
1.4	B	F	57

- **Gradient Boost, Step 1**
  - Leaf의 계산
  - Target인 Weight의 평균: 71.2
  - Residual을 계산: 실제값과 예측값의 차이(error)

같은 X변수들로
Residual에 대한 Tree

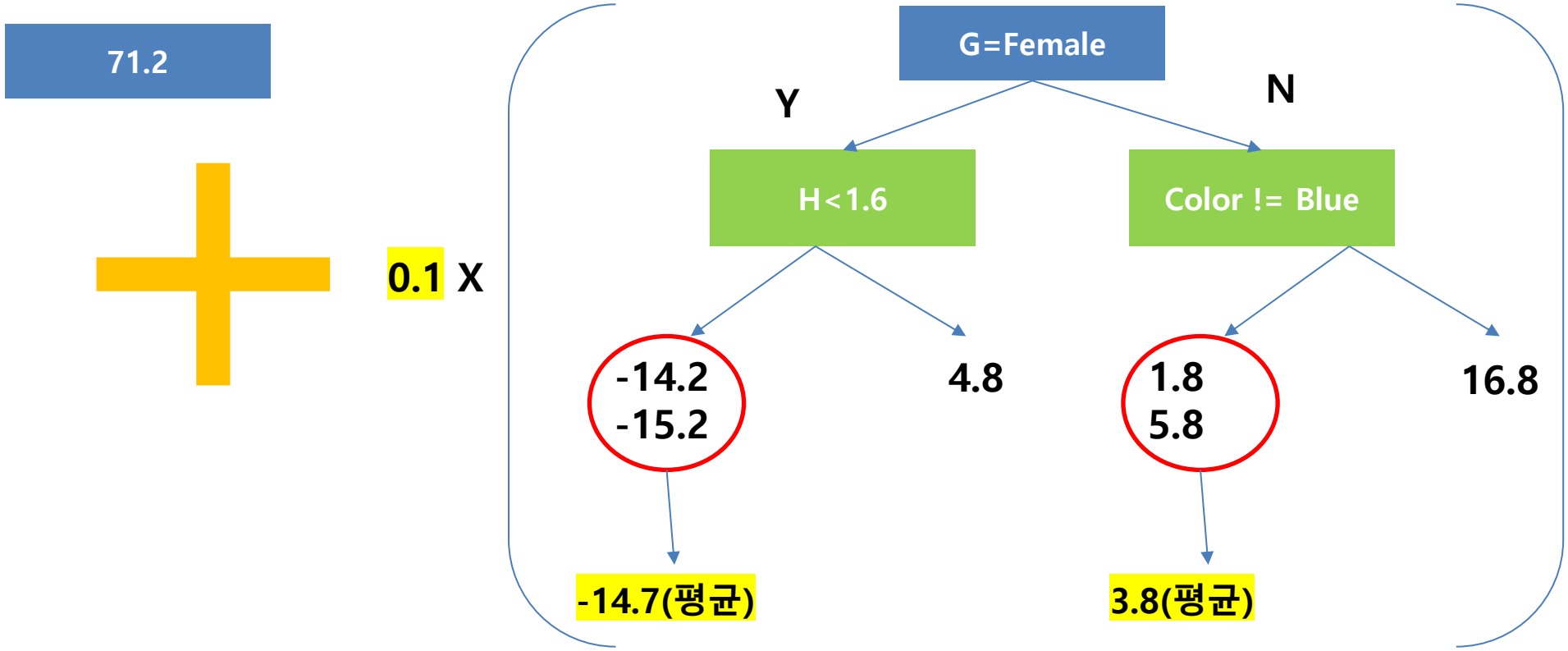
Height	Color	Gender	Weight	Residual
1.6	B	M	88	16.8
1.6	G	F	76	4.8
1.5	B	F	56	-15.2
1.8	R	M	73	1.8
1.5	G	M	77	5.8
1.4	B	F	57	-14.2

- **Gradient Boost, Step 1**
  - **Leaf + 1st Tree**



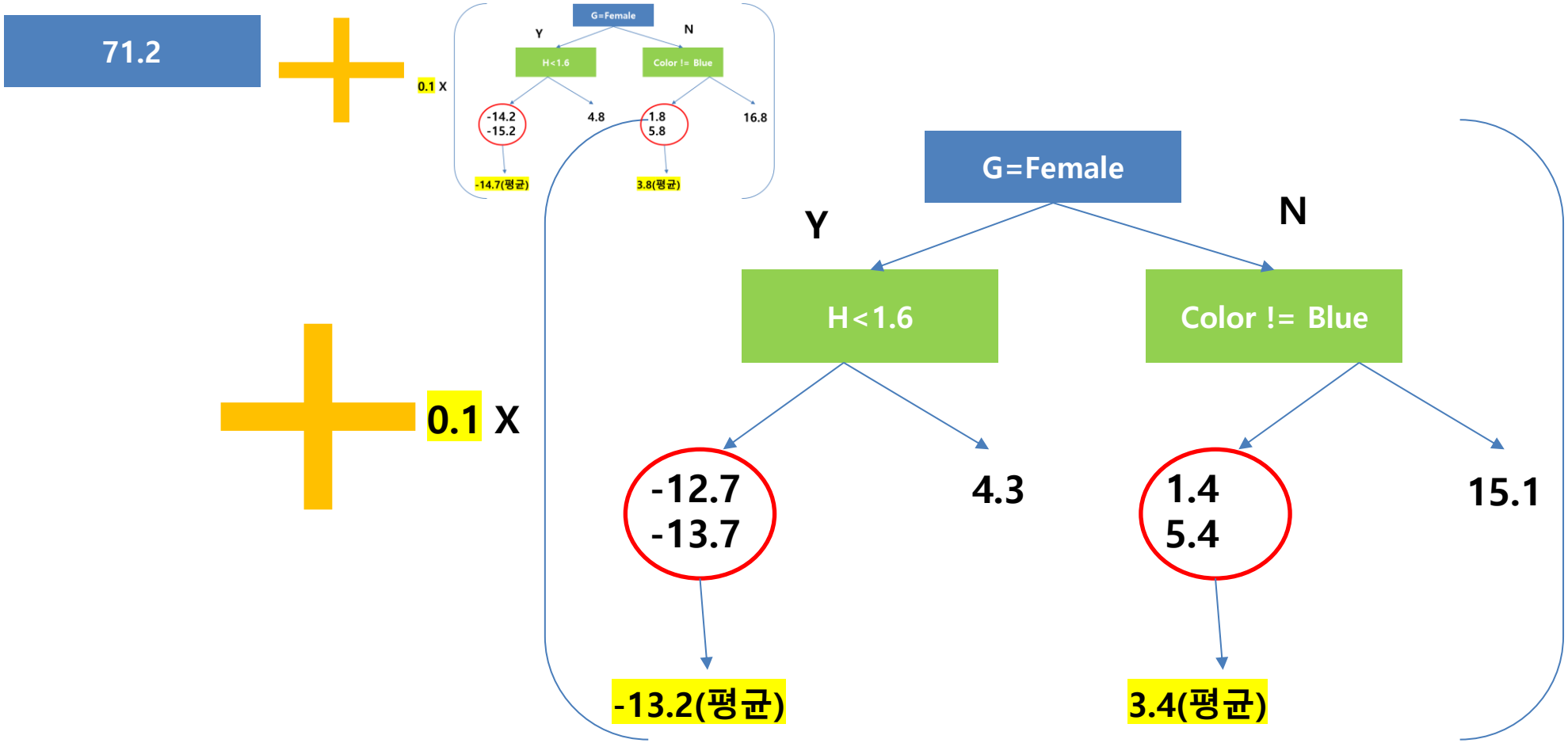
- **Male, Blue인 경우 예측 예시:**
  - 71.2 + 16.8 = 88 (관측치와 동일하지만 과적합)
  - Bias는 작지만 Variance 큰 상태

- **Gradient Boost, Step 2**
  - 과적합 방지, 학습속도 조절을 위한 학습율 도입
  - Learning Rate: 0~1사이, 이 예에서는 0.1 사용



- **Male, Blue인 경우 예측 예시:** 71.2 + 0.1 X 16.8 = 72.9
  - 실제값에 가까워지지만, 그 정도가 조절됨 (Gradient의 개념)
  - Variance를 낮게 유지할 수 있음

- **Gradient Boost, Step 3**
  - Learning Rate: 0~1사이, 이 예에서는 0.1 사용



- **H=1.6, Male, Blue인 경우 예측 예시:** 71.2 + 0.1 X 16.8 + 0.1 X 15.1 = 74.4

- **Gradient Boost, Step 3**
  - 학습율 반영 예측값을 통한 두 번째 Residual 계산

같은 X변수들로 New Residual에 대한 Tree

Height	Color	Gender	Weight	Residual	Residual(new)
1.6	B	M	88	16.8	15.1
1.6	G	F	76	4.8	4.3
1.5	B	F	56	-15.2	-13.7
1.8	R	M	73	1.8	1.4
1.5	G	M	77	5.8	5.4
1.4	B	F	57	-14.2	-12.7

Residual 크기 감소

- **Gradient Boost**
  - 위의 과정을 계속 반복
    - 정해진 iteration한도 까지 반복
    - 또는 이전 단계와 이후 단계의 Residual 차이가 없을 때까지 반복

  - 매 iteration에서의 Tree의 leaf는 8~32개 사이에서 생성
  - 매 iteration마다 다르게 생성
    - 1st tree: leaf 8개
    - 2nd tree: leaf 32개
    - 3rd tree: leaf 16개
    - …

Industrial Data Science Lab

Contact:
won.sang.l@gwnu.ac.kr
https://sites.google.com/view/idslab