



HD현대인프라코어 재직자 2과정 보충자료

회귀 분석

01 예측 (Prediction)

예측 (Prediction)

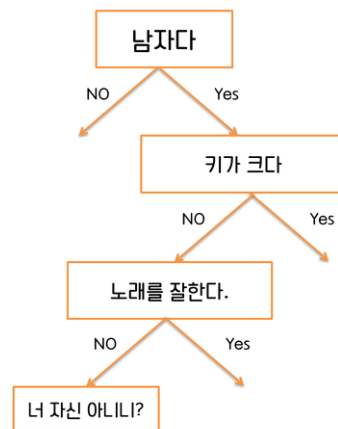
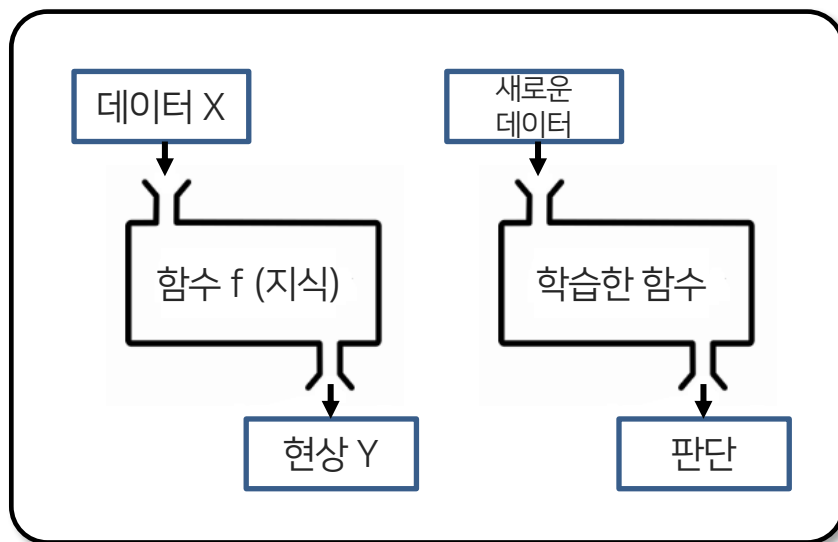
- 예측은 과거 데이터와 과거 패턴 연구를 기반으로 한 예상

-> 머신 러닝, 통계 모델링, 데이터 마이닝과 같은 분석 기술을 사용하여 조직이 트렌드, 행동, 향후 성과, 비즈니스 기회 등을 파악할 수 있도록 지원

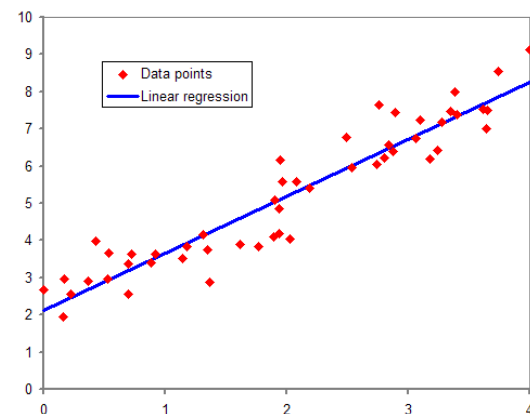


❖ 머신러닝 (Machine Learning)

특정 알고리즘을 통해 데이터를 분석하고 분석 결과를 스스로 학습한 후,
이를 기반으로 판단을 진행하는 것

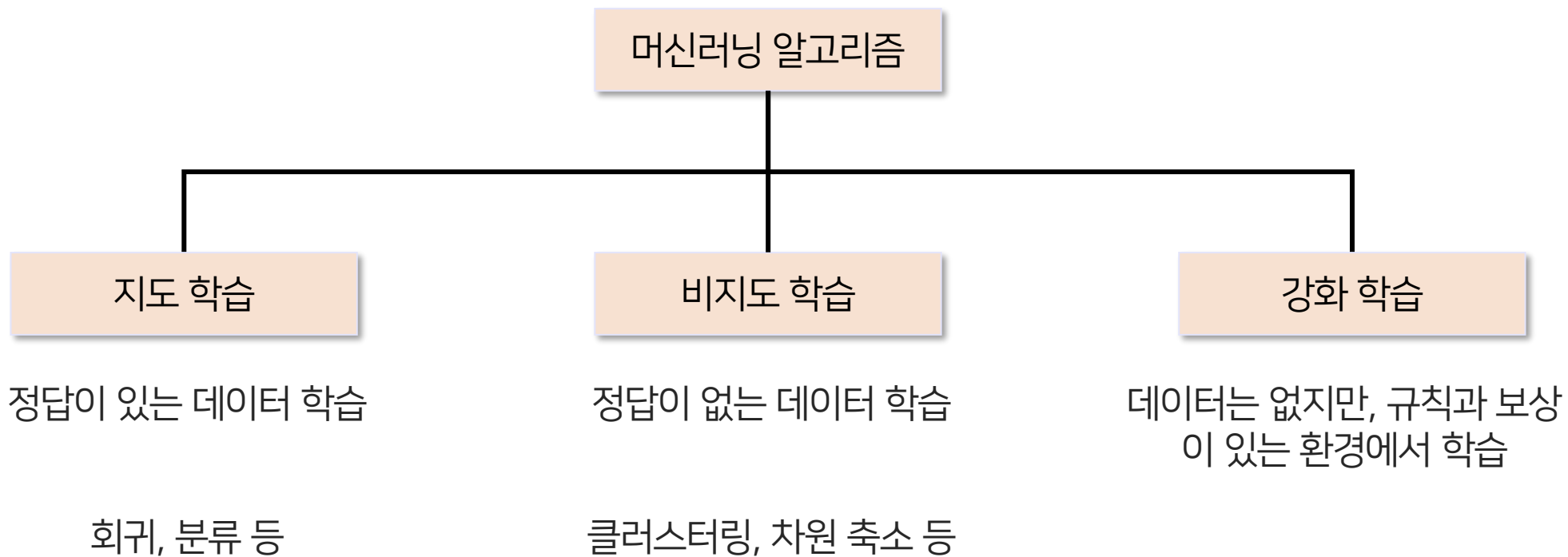


의사결정나무
(Decision Tree)

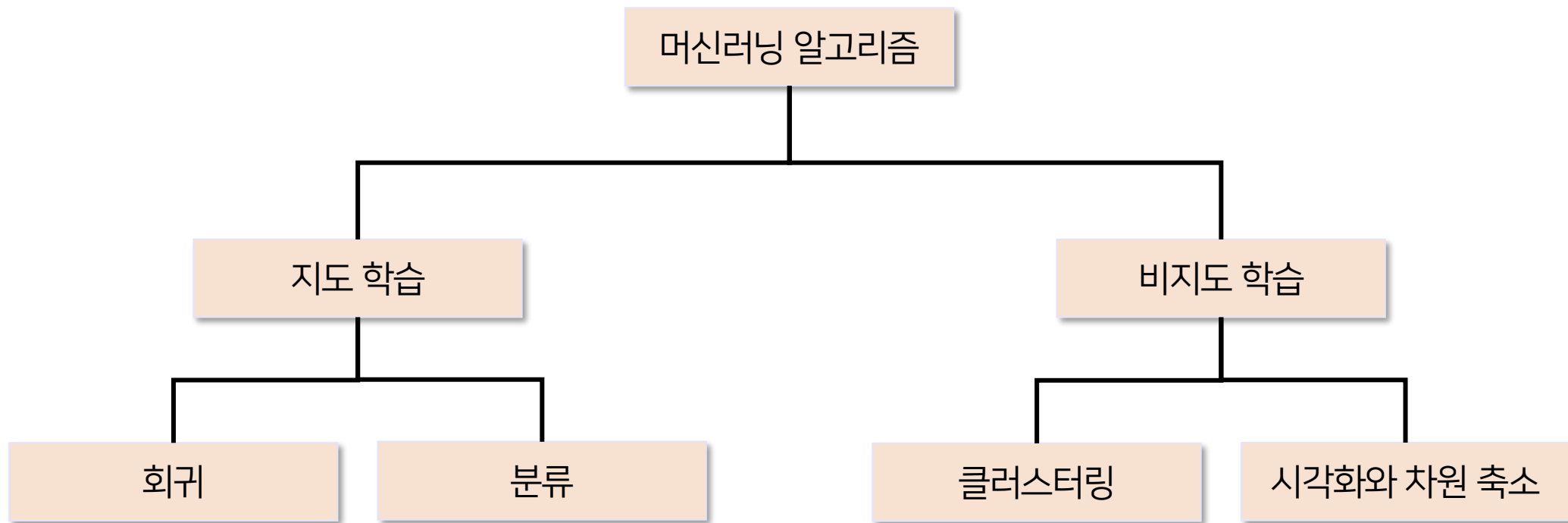


선형회귀
(Linear Regression)

◆ 머신러닝의 분류



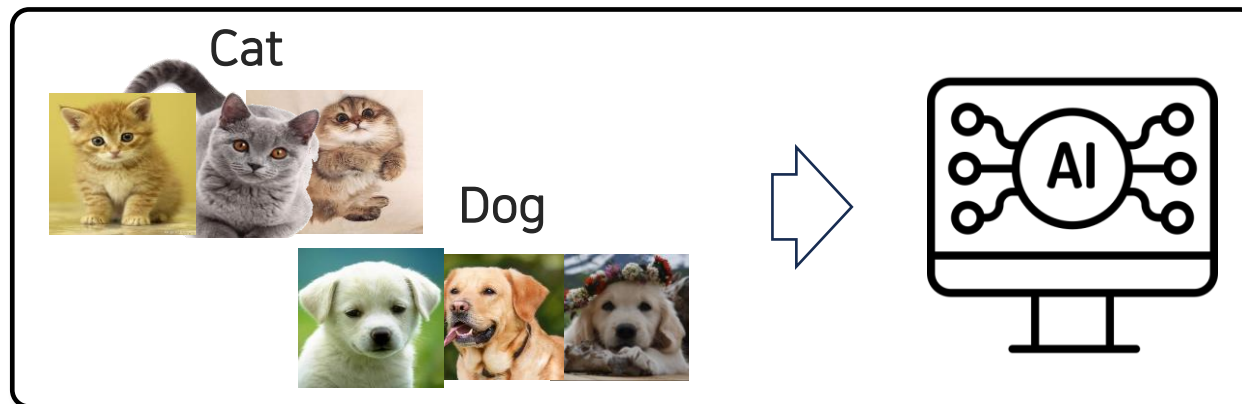
❖ 머신러닝 (Machine Learning)



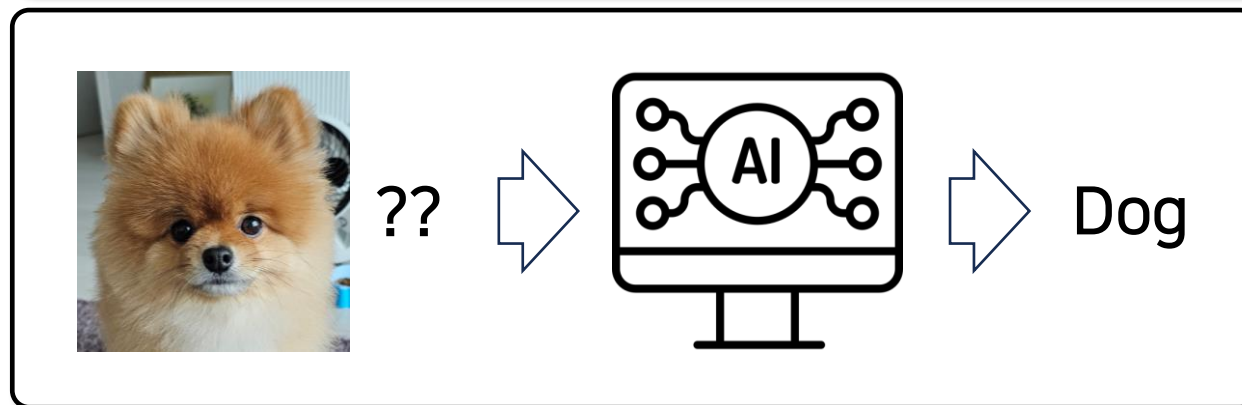
❖ 머신러닝 알고리즘 - 지도학습 (Supervised Learning)

정답이 있는 훈련 데이터를 사용하여 학습을 진행하는 알고리즘

Train 과정

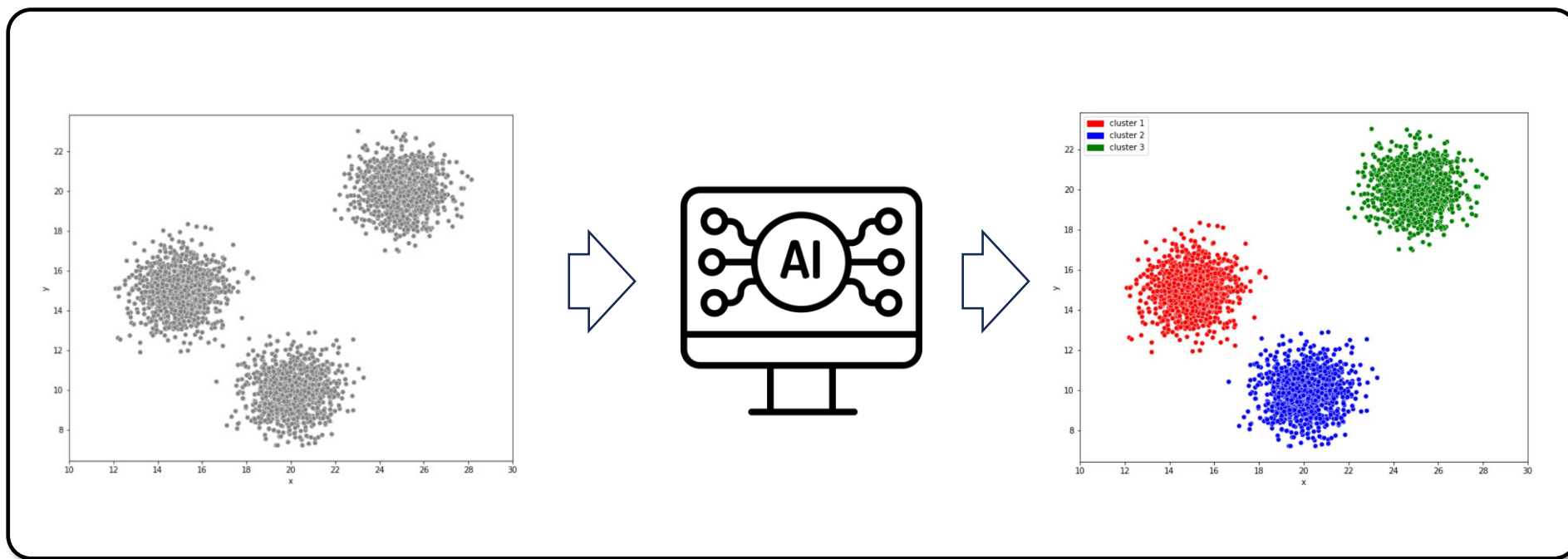


Test 과정



❖ 머신러닝 알고리즘 - 비지도학습 (Unsupervised Learning)

정답이 없는 훈련 데이터를 사용하여 학습을 진행하는 알고리즘



회귀 분석 (Regression Analysis)

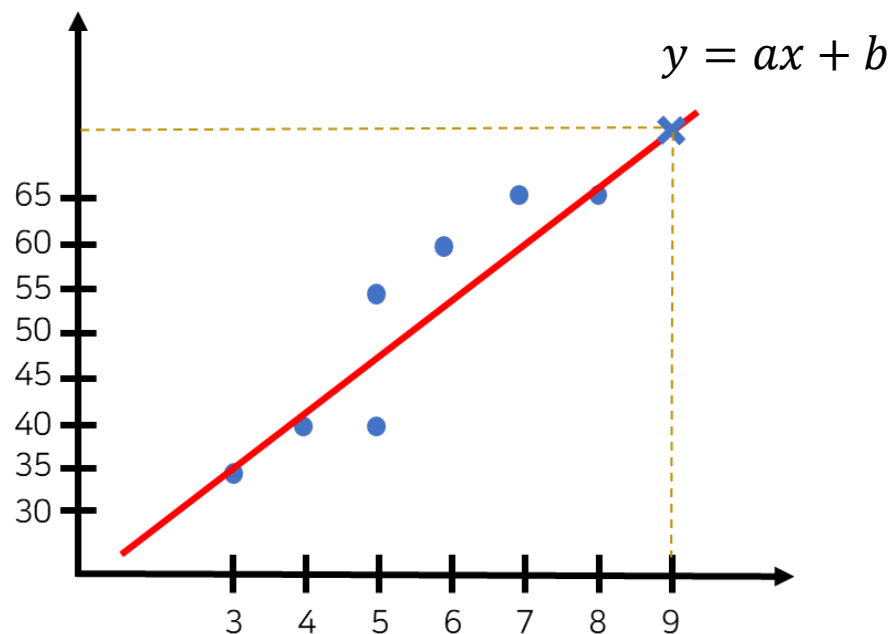
다음 표는 A 중학교 학생들의 영어 공부 시간과 공부시간에 따른 영어 시험 성적을 나타내고 있다.
만약 철수가 9시간동안 공부를 한다면 성적은 몇 점을 받을 수 있을까?

공부시간	시험 성적
3시간	30점
3시간	35점
4시간	40점
5시간	55점
5시간	40점
6시간	60점
7시간	65점
9시간	?

회귀 분석 (Regression Analysis)

- 2개의 변수 사이의 관계를 분석하는 방법 중 하나
- 회귀: $y = f(x)$ 라는 함수를 통해 변수 사이의 관계를 공식화하는 것

공부시간	시험 성적
3시간	35점
4시간	40점
5시간	55점
5시간	40점
6시간	60점
7시간	65점
8시간	65점
9시간	?



◆ 단순 선형회귀 (Simple Linear Regression)

- 회귀분석을 사용하는 회귀식이 '파라미터에 관한' 1차식이 되는 경우
- 기본 회귀식: $y = ax + b + \varepsilon$

$$y_i = a_i x + b_i + \varepsilon_i \quad (y_i(y_1, y_2, \dots, y_n), x_i(x_1, x_2, \dots, x_n))$$

➔ 설명변수 x 에 대해 $ax + b$ 를 계산한 값에, 확률오차 ε 를 더한 값

➔ y 는 반응 변수, a 와 b 는 회귀계수가 됨 (Regression Coefficient)

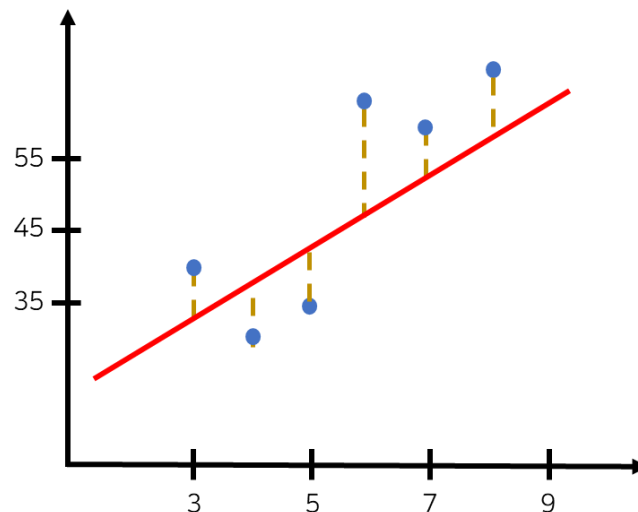
회귀 분석의 목표는 **아직 미지수인 회귀계수를 적절한 값으로 추정하는 것!**

단순 선형회귀 (Simple Linear Regression)

- 가장 기초적이나 많이 사용되는 알고리즘
- 입력값(X)가 1개일 경우에만 적용 가능
- 입력값과 결과값의 관계를 알아보기 가장 용이함
- 입력값이 결과값에 얼마나 영향을 미치는 지 알 수 있음
- 두 변수 간의 관계를 직관적으로 해석하고자 하는 경우 활용

◆ 단순 선형회귀 (Simple Linear Regression)

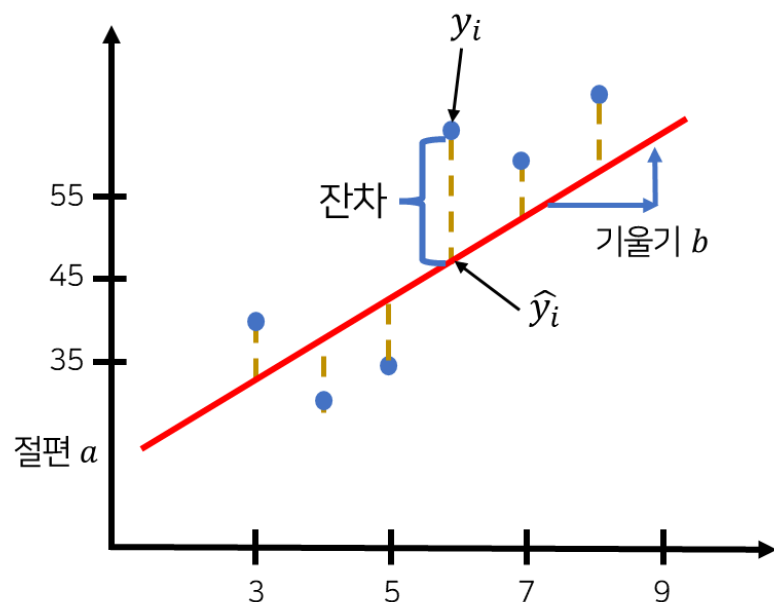
회귀 분석의 목표는 아직 미지수인 회귀계수를 적절한 값으로 추정하는 것!



즉, 데이터와 회귀식의 차이가 최대한 작은 회귀식을 찾는 것이 목표

❖ 최소제곱법 (OLS, Ordinary Least Squares)

- 실제 값과 회귀 값 사이의 잔차를 최소화하는 방식을 통해 모회귀계수를 측정하는 방법
- 잔차(Residual): 회귀식과 관측값 사이에 나타나는 차이. $\hat{y}_i - y_i$
- 잔차 제곱의 총합 $E(a, b)$ 를 최소화하는 a, b 를 찾기



[모형에서 얻은 y]

$$\hat{y}_i = a + bx_i$$

[실제 y]

y_i

[잔차 제곱의 총합]

$$E(a, b) = \sum_{i=1}^n (\hat{y}_i - y_i)^2 = \sum_{i=1}^n (a + bx_i - y_i)^2$$

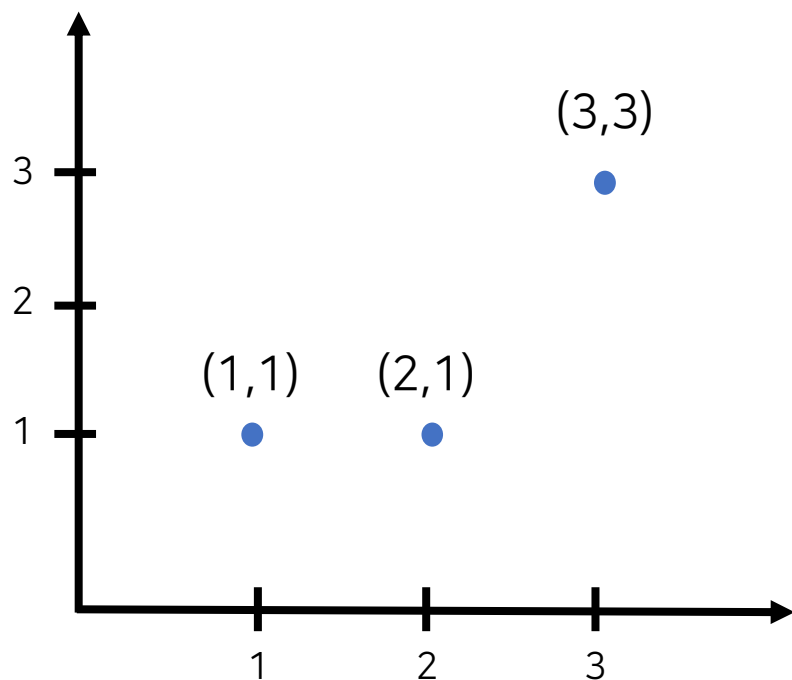
$E(a, b)$ 가 최소가 되는 a, b 를 구하려면,

$$\frac{\partial}{\partial a} E(a, b) = 0, \quad \frac{\partial}{\partial b} E(a, b) = 0$$

를 만족하는 a 와 b 를 구한다.

❖ 최소제곱법 (OLS, Ordinary Least Squares)

위 세 개의 점을 적합하는 회귀식의 회귀계수 a 와 b 를 구해 봅시다.



◈ (참고) 선형 회귀 계수 유도

$$\frac{\partial L}{\partial \beta_0} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial L}{\partial \beta_1} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) x_i = 0$$

$$\sum_{i=1}^n y_i - \beta_1 \sum_{i=1}^n x_i - n\beta_0 = 0 \quad \longrightarrow \quad \beta_0 = \frac{1}{n} \sum_{i=1}^n y_i - \beta_1 \frac{1}{n} \sum_{i=1}^n x_i \quad \longrightarrow \quad \therefore \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\sum_{i=1}^n x_i y_i - \beta_0 \sum_{i=1}^n x_i - \beta_1 \sum_{i=1}^n x_i^2 = 0$$

$$\Rightarrow \sum_{i=1}^n x_i y_i - \left(\frac{1}{n} \sum_{i=1}^n y_i - \beta_1 \frac{1}{n} \sum_{i=1}^n x_i \right) \sum_{i=1}^n x_i - \beta_1 \sum_{i=1}^n x_i^2 = 0$$

$$\Rightarrow \sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n y_i \sum_{i=1}^n x_i + \beta_1 \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 - \beta_1 \sum_{i=1}^n x_i^2 = 0$$

$$\Rightarrow \beta_1 \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 - \beta_1 \sum_{i=1}^n x_i^2 = \frac{1}{n} \sum_{i=1}^n y_i \sum_{i=1}^n x_i - \sum_{i=1}^n x_i y_i$$

$$\therefore \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

❖ 최소제곱법 - 예측선 그리기

- 회귀모형 $y = ax + b$ 의 회귀계수 a 와 b 를 결정하는 방법
- $a = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$, $b = \bar{y} - a * \bar{x}$

[예제]

공부시간	시험 성적	예측 값
3시간	35점	36.71
5시간	55점	51.86
6시간	60점	59.43
7시간	65점	67
9시간	?	82.14

$$\bar{x} = \frac{(3 + 5 + 6 + 7)}{4} = \frac{21}{4} = 5.25$$

$$\bar{y} = \frac{(35 + 55 + 60 + 65)}{4} = 53.75$$

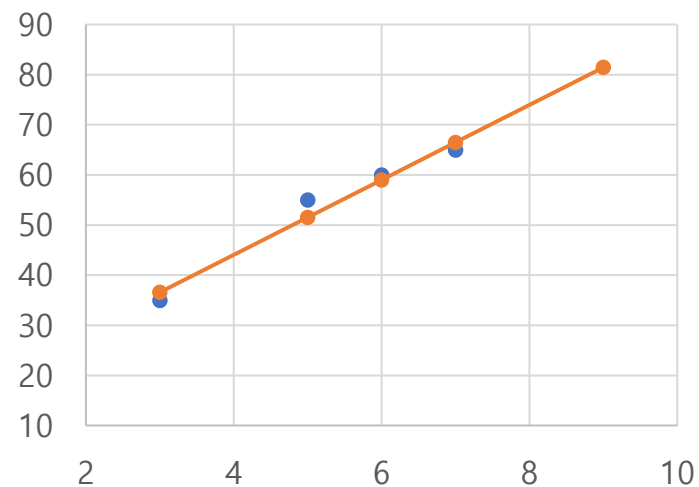
$$a = \frac{(3-5.25)(35-53.75)+(5-5.25)(55-53.75)+(6-5.25)(60-53.75)+(7-5.25)(65-53.75))}{(3-5.25)^2+(5-5.25)^2+(6-5.25)^2+(7-5.25)^2} = 7.57$$

$$b = 53.75 - 7.57 * 5.25 = 14$$

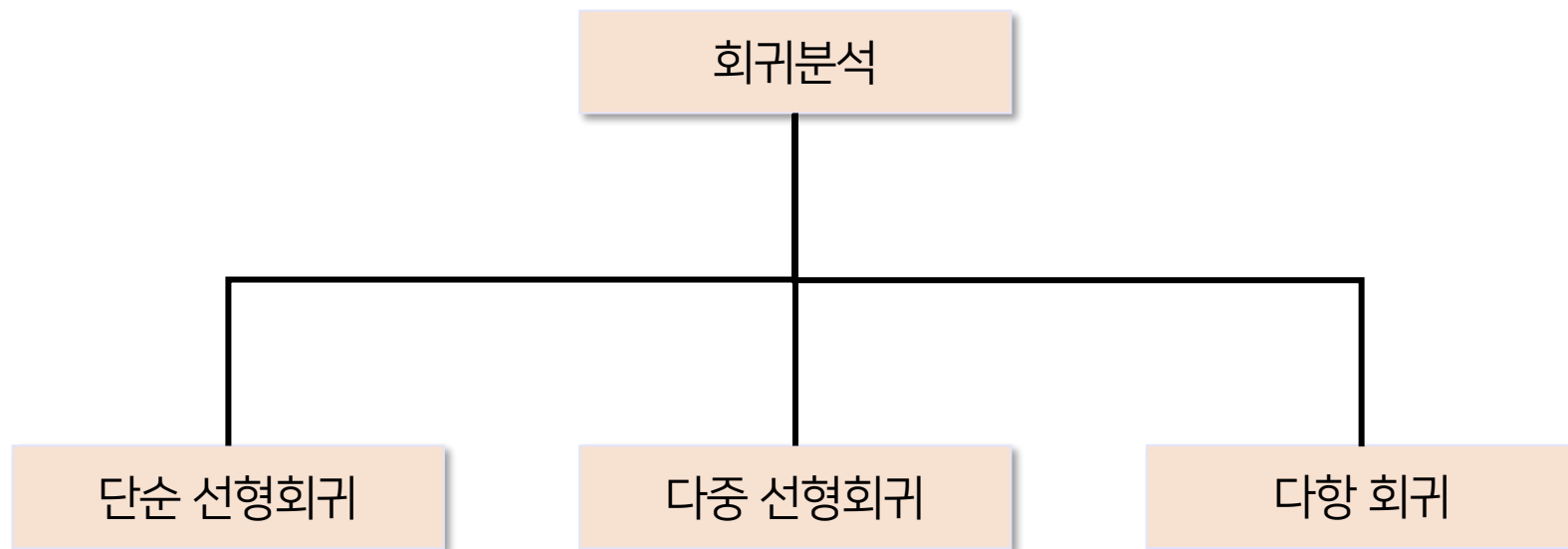
$$y = 7.57x + 14$$

❖ 최소제곱법 - 예측선 그리기

공부시간	시험 성적	예측 값
3시간	35점	36.71
5시간	55점	51.86
6시간	60점	59.43
7시간	65점	67
9시간	?	82.14

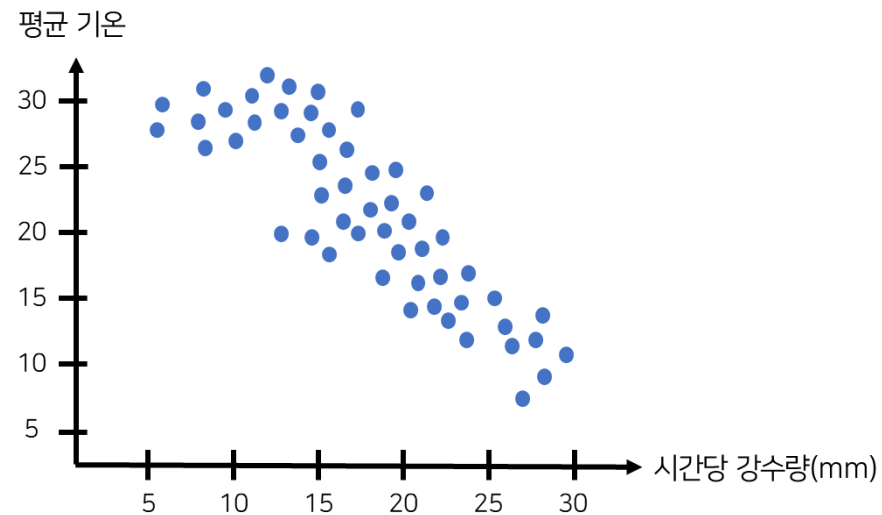
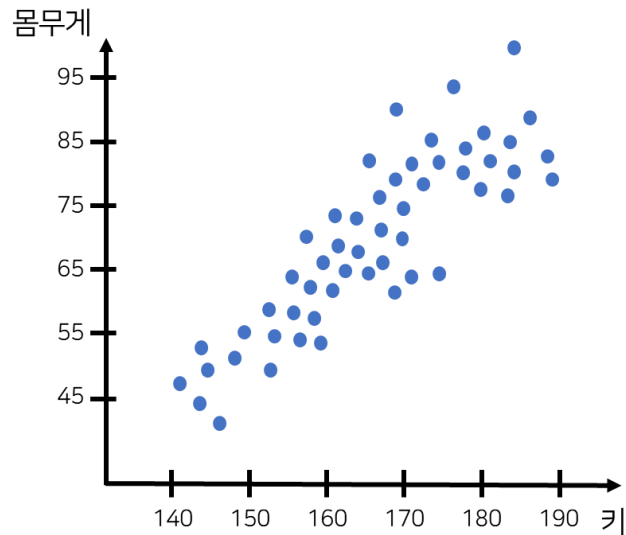


◈ 선형회귀 (Linear Regression)



◆ 단순 선형회귀

- 가장 기본적이고 단순한 회귀 모형
- $y = ax + b$
- 입력 값 X와 출력 값 Y 사이의 관계를 이해할 때 주로 사용
- 입력 값이 1개인 경우에 사용



◆ 다중 선형회귀

- $y = a_1x_1 + a_2x_2 + \cdots + a_nx_n + b$
- 입력 값이 2개 이상인 경우에 사용
- 개별 x_i 에 대해 적절한 a_i 를 찾아야 함

공부시간 (x_1)	수행 평가 (x_2)	시험 성적 (y)
3시간	5점	35점
5시간	7점	55점
6시간	10점	60점
7시간	13점	65점
9시간	13점	?

방 수 (x_1)	층 수 (x_2)	화장실 수 (x_2)	건설 년도 (x_2)	주택 가격 (y)
1	1	1	2004	10,000
2	1	2	2010	100,000
2	1	2	2005	50,000
4	2	3	2017	999,000
3	2	1	2013	?

다중 선형회귀

- 여러 개의 입력값과 결과값 간의 관계 확인 가능
- 어떤 입력값이 결과값에 어떠한 영향을 미치는지 알 수 있음.
- 회귀모형이 복잡해짐에 따라 편의상 벡터와 행렬을 많이 사용함
- 여러 개의 입력값 사이 간의 상관 관계가 높을 경우 결과에 대한 신뢰성을 잃을 가능성이 있음

◆ 다중 선형회귀

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i, \quad i = 1, 2, \dots, n$$

$$i = 1 \text{인 경우, } y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_{p-1} x_{1,p-1} + \epsilon_1$$

$$i = 2 \text{인 경우, } y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_{p-1} x_{2,p-1} + \epsilon_2$$

...

$$i = n \text{인 경우, } y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_{p-1} x_{n,p-1} + \epsilon_n$$



$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ 1 & x_{31} & x_{32} & \cdots & x_{3,p-1} \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Y

X

최소제곱법을
이용한 추정



$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_{p-1} \end{pmatrix} = (X^T X)^{-1} X^T Y$$

◈ 변수선택법

- 변수가 여러 개 있는 모형에서 최선의 변수 조합을 찾아내는 기법
- 전진선택법, 후진제거법, 단계별선택법 존재
- 변수를 추가하거나 제거하며 AIC (Akaike Information Criterion)가 가장 낮은 모형 채택 ($AIC = 2k - 2 \ln(L)$, k 는 매개변수 개수, L 은 우도)

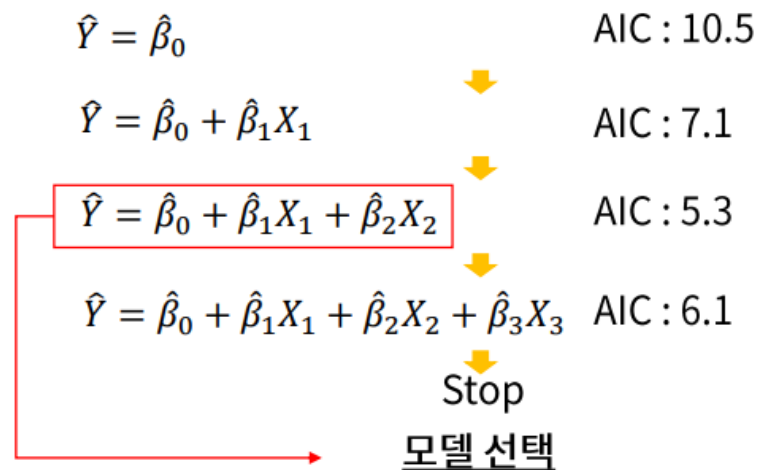
전진선택법
(Forward
Selection
Procedure)

후진제거법
(Backward
Elimination
Procedure)

단계적 선택법
(Stepwise
Method)

◆ 전진선택법 (Forward Selection Procedure)

- 변수 하나부터 개수를 추가하며 성능지표를 비교하는 방법
- 예시) 영모형 $Y = 1$ 에서 변수를 하나씩 추가해가면서 모형을 선택
- 중요하다고 생각되는 변수부터 차례로 모형에 추가하면서 설명력을 확인
- 추가하면서 AIC의 변화를 확인하며 변수 추가를 멈추는 식으로 사용



◆ 후진제거법 (Forward Selection Procedure)

- 모든 변수를 포함한 모형에서 하나씩 제거하면서 비교하는 방법
- 예시) 전체 모형에서 하나씩 변수를 제거해가면서 모형을 선택
- 가장 적은 영향을 주는 변수부터 하나씩 제거하면서 더 이상 제거할 수 없을 때까지 진행

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4 \quad \text{AIC : 5.5}$$



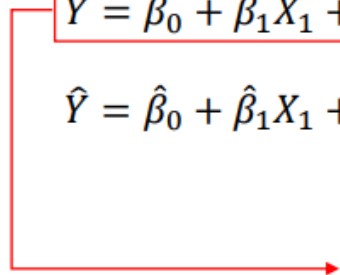
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 \quad \text{AIC : 5.1}$$



$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 \quad \text{AIC : 5.3}$$

Stop

모델 선택



◆ 단계적 선택법 (Stepwise Method)

- 모든 부분집합을 고려하는 방법으로 최적의 변수를 선택 가능
- 전진선택법에 따라 변수를 추가한 후, 기존 변수의 중요도가 낮아질 시 해당 변수를 제거하는 등 추가 또는 제거되는 변수 여부를 판단, 더 이상 없을 때에 중단

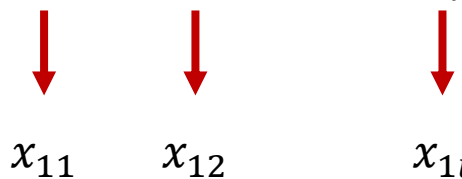
❖ 다중공선성 (Multicollinearity)

- 독립변수 간에 강한 상관관계가 존재하는 경우
- 다중공선성이 존재하면 회귀모델의 성능을 평가하거나 독립변수들의 영향력을 해석하는 것이 어려워질 수 있음
- 주로 상관계수나 분산팽창계수 (VIF, Variation Inflation Factor)를 통해 평가됨
- 일반적으로 VIF가 10 이상일 때 다중공선성이 존재한다고 판단함 * $VIF : \frac{1}{1-R_i^2}$
- 변수선택법 적용 시 다중공선성이 존재하는 독립변수들이 있으면 일반적으로 해당하는 변수들 중 하나만 남기고 나머지는 제거한 후 분석

◆ 다항회귀 (Polynomial Regression)

- 1차 함수 선형식으로 표현하기 어려운 분포의 데이터를 위한 회귀
- 복잡한 분포의 데이터의 경우 일반 선형 회귀 알고리즘 적용 시 낮은 성능의 결과가 도출됨
- 데이터의 분포에 더 잘 맞는 모델이 필요함
- 다항회귀의 거듭제곱인 독립변수들을 다른 변수로 대체하면 다중회귀와 같음
- 다항회귀는 독립변수들을 좀더 복잡한 값으로 만들어 선형회귀에 넣어 학습하는 것

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \cdots + \beta_i x_1^i$$


$$x_{11} \quad x_{12} \quad x_{1i}$$

치환하면 결국 다중회귀와 같음

회귀식의 정도(precision)

- 추정된 회귀식이 원래의 관측값들을 어느 만큼 대표하는지를 나타내는 척도
- MSE, MAE, 결정계수 등이 많이 사용됨

MSE (Mean Square Error) : $\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2$

- 이상치에 대해 민감하다. 즉, 정답에 대해 예측값이 매우 다른 경우, 그 차이는 오차값에 상대적으로 크게 반영된다.

MAE (Mean Absolute Error) : $\frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i|$

- MSE보다 이상치에 둔감 혹은 강건함.

RMSE (Root Mean Square Error) : $\sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}$

- MSE에 루트를 씌워 사용. 상대적으로 비직관적이지만 MAE보다 이상치에 대해 강건

◆ 결정계수 (Coefficient of determination)

y_i : 관측값

\bar{y} : 관측값의 평균

\hat{y} : 회귀식의 예측값

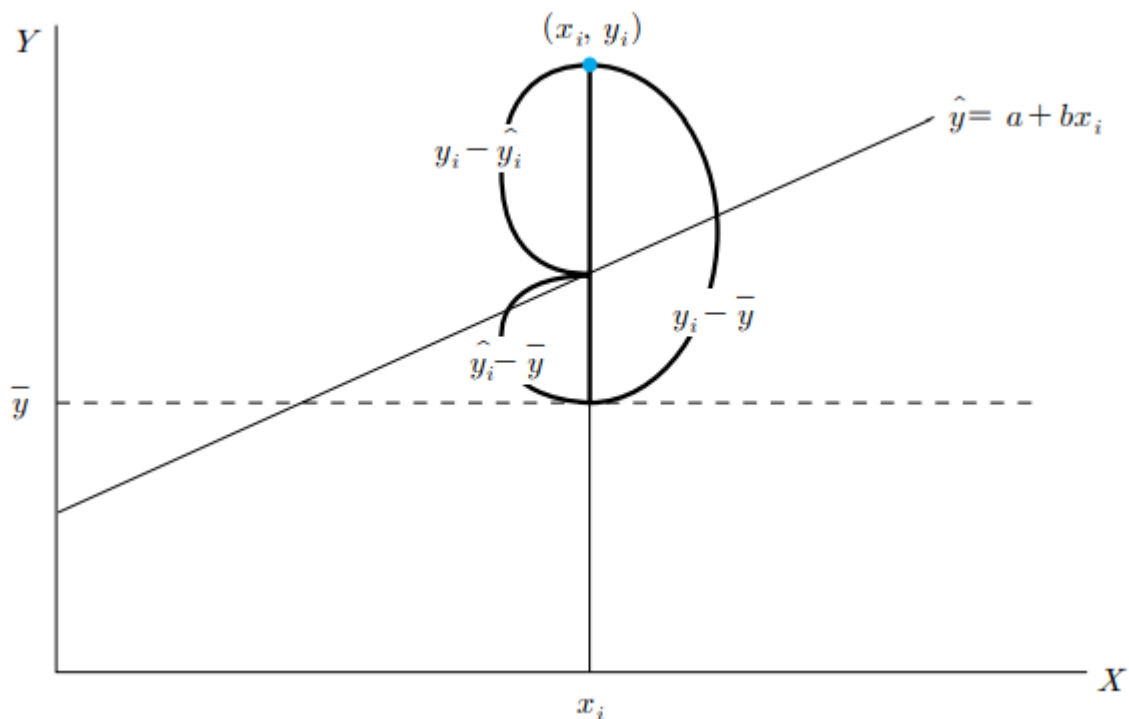
관측값 y_i 의 편차 $y_i - \bar{y}$ 를 분해하면

$$\underbrace{y_i - \bar{y}}_{\text{①}} = \underbrace{(y_i - \hat{y}_i)}_{\text{②}} + \underbrace{(\hat{y}_i - \bar{y})}_{\text{③}}$$

① 편차

② 회귀식에 의해 설명되지 않는 편차

③ 회귀식에 의해 설명되는 편차



◆ 결정계수 (Coefficient of determination)

$$\frac{y_i - \bar{y}}{\text{①}} = \frac{(y_i - \hat{y}_i)}{\text{②}} + \frac{(\hat{y}_i - \bar{y})}{\text{③}}$$

이 식을 모든 편차들에 대해 제공한 후 합하면

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

$SST = SSE + SSR$

(총제곱합) (잔차제곱합) (회귀제곱합)

SSE가 작을수록, SSR이 클수록 회귀적합이 잘 된 것임

결정계수 : 총제곱합 SST 중에서 회귀제곱합 SSR이 차지하는 비율

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

* 결정계수 = (상관계수)²

◆ 결정계수 (Coefficient of determination)

위 세 개의 선을 적합하는 회귀식의 결정계수를 구해 봅시다.

