CSFSM7053-00
Business Analytics

College of Hotel & Tourism
Kyung Hee University

Lecture Note 07
선형 회귀분석 (Linear Regression)

Dohyung Bang

Fall, 2021

Syllabus

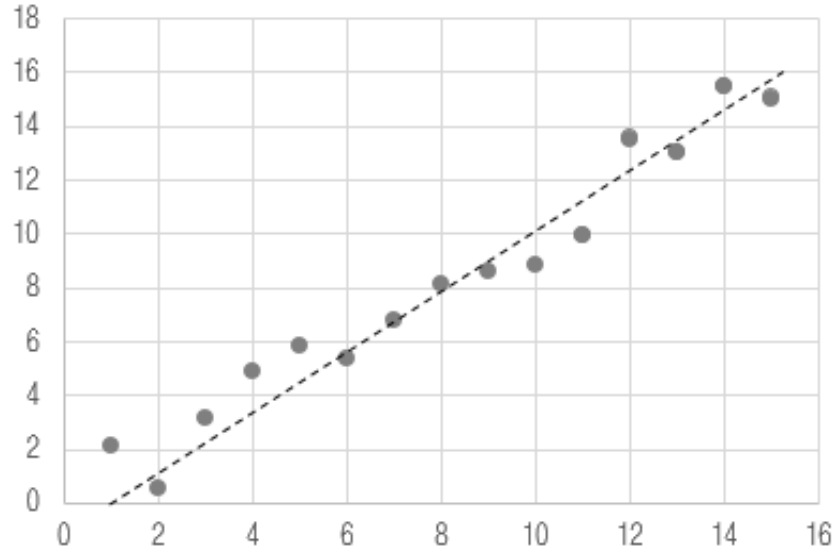
Week	Date	Topic	Note
1	9/6(월)	R Basic - R 기초 문법 학습	
2	9/13(월)	R Basic – Data Manipulation I	과제#1
3	9/20(월) (추석)	<추석> (보충영상) R Basic - Data Manipulation II	
4	9/27(월)	Descriptive Analytics I - 데이터 요약하기/상관관계/차이검증	~~과제#2~~
5	10/4(월) (대체공휴일)	<대체공휴일> (보충영상) Descriptive Analytics II - 데이터 시각화	과제#2
6	10/11(월) (대체공휴일)	<대체공휴일> (보충영상) Supplementary Topic I - 외부 데이터 수집 (정적 컨텐츠 수집)	~~과제#4~~ 과제#3
7	**10/18(월)**	**Predictive Analytics I – Linear regression**	
8	10/25(월)	Predictive Analytics II – Logistic Regression	시험 대체 수업
9	11/1(월)	Predictive Analytics III – Clustering & Latent Class Analysis	과제#4
10	11/8(월)	Predictive Analytics IV – Tree-based Model and Bagging (Random Forest)	
11	11/15(월)	Predictive Analytics V – Association Rules	
12	11/22(월)	Prescriptive Analytics I – Linear Programming	과제#5
13	11/29(월)	Prescriptive Analytics II – Data Envelopment Analysis (DEA)	
14	12/6(월)	Prescriptive Analytics III – Integer Programming	과제#6
15	12/13(월)	Prescriptive Analytics IV – Simulation	Quiz
16	**12/20(월)**	**Final Presentation**	

Lecture 7-1

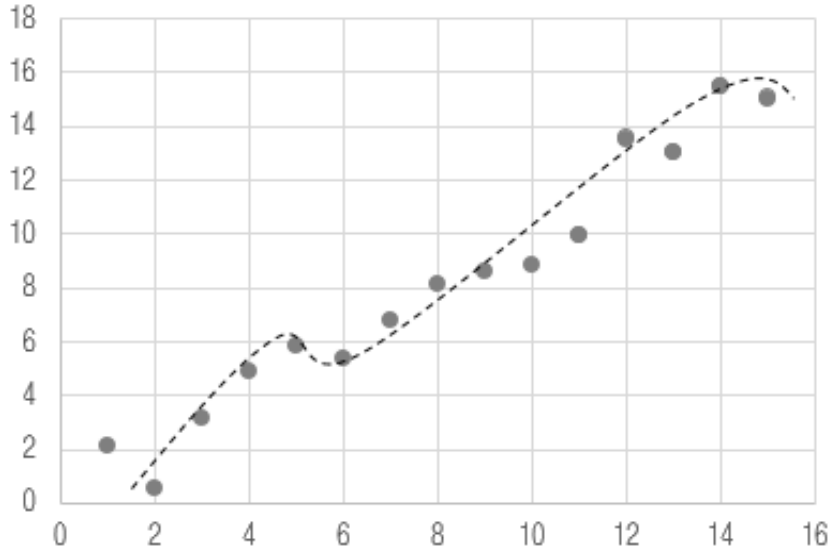
회귀분석 이해

회귀분석(Regression)이란 무엇인가?

좀 더 데이터 사이언스 관점에서 정의하면, Input 변수인 X를 이용해 Output 변수인 Y를 예측하는 방법으로, 모형의 가정에 따라 선형회귀모형과 비선형 회귀모형으로 나눌 수 있음



선형 회귀 모형



비선형 회귀 모형

회귀분석의 적용 및 응용

"광고가 매출액에 얼마나 영향을 미치는가?"

"제품판매 가격을 어떻게 결정할 것인가?"

"매장을 오픈했을 때, 예상되는 매출은 얼마일까?"

"소비자 수요가 얼마나 될까?"

⋮

"비만이 성인병 발생에 미치는 영향은 얼마나 될까?"

"세포활동이 암 발병에 미치는 영향은 어떻게 될까?"

"신약의 성분변화에 따라 치료율이 얼마나 될까?"

⋮

"교육수준에 따라 소득이 어떻게 달라질까?"

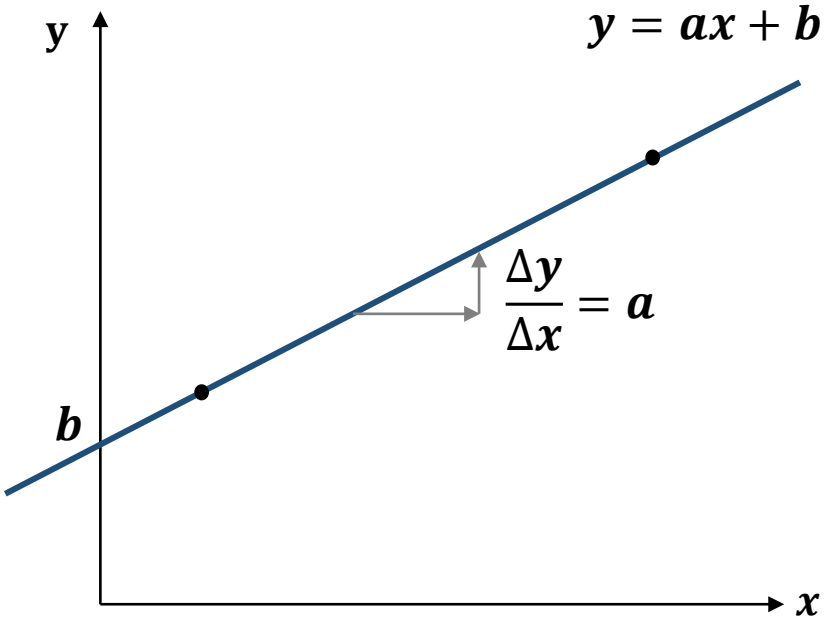
" 조직성과가 직무만족에 미치는 영향은 얼마나 될까?"

⋮

회귀분석은 인과관계의 정도를 정량적으로 분석하고자 하는 여러 맥락에서 적용가능함

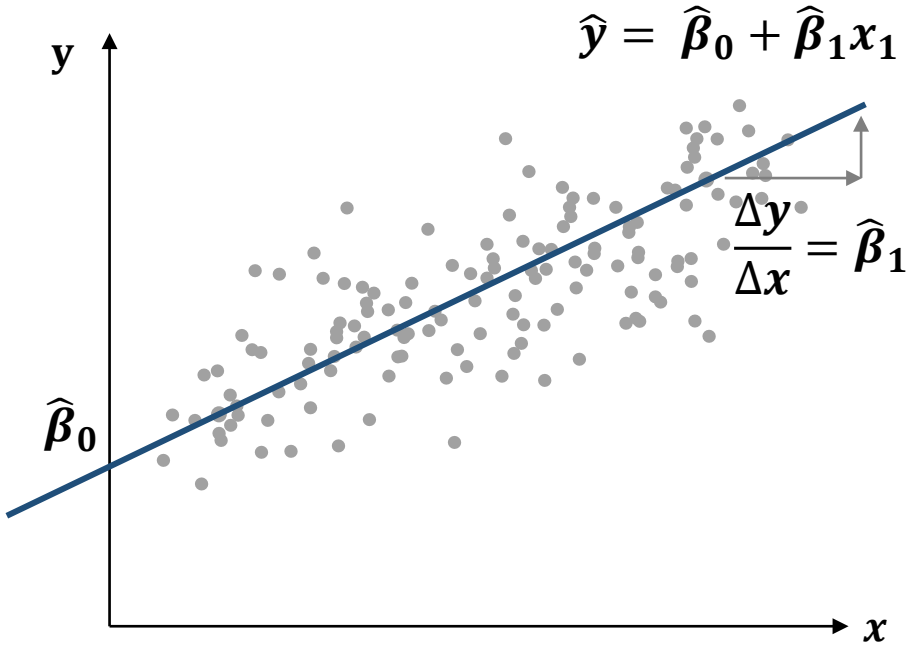
선형 회귀분석과 1차 방정식

우리가 배운 선형방정식을 Remind 해보자.



- ✓ **기울기(Slope)**와 **절편(Intercept)**이 주어지면 1차방정식 즉, 1차 함수 (선형함수) 형태로 나타낼 수 있음
- ✓ 1차 방정식에서는 기울기와 절편은 주어지는 값

회귀분석 = 1차방정식을 찾는 과정



- ✓ **회귀분석은 1차 함수로 관계를 가장 잘 나타내기 위해 데이터로부터 역으로 기울기(Slope)와 절편(Intercept)를 찾아가는 과정**
- ✓ **기울기의 크기에 따라 두 변수 간 인과관계의 민감도를 판단할 수 있음**

선형 회귀분석(Linear Regression)

True Model

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

내가 가정한 관계에서 회귀모형으로 설명되는 부분

오차항(Error term)
: 회귀모형으로 설명되지 않는 부분

- 우리가 알고 싶은 현상을 표현한 식
- **진리; 가상의 모형; 오직 신(God)만 알고 있음**
- 우리가 모르는 모집단에서의 'x, y **관계가 이럴 것이다**' 가정한 식.
- $\alpha, \beta, \varepsilon$**는** 모르지만 모집단의 관계를 설명해주는 모수(Parameter)로 추정의 대상이 됨.

My Model

$$y_i = \widehat{\beta_0} + \widehat{\beta_1} x_i + e_i = \hat{y}_i + e_i$$

$\hat{y}_i$

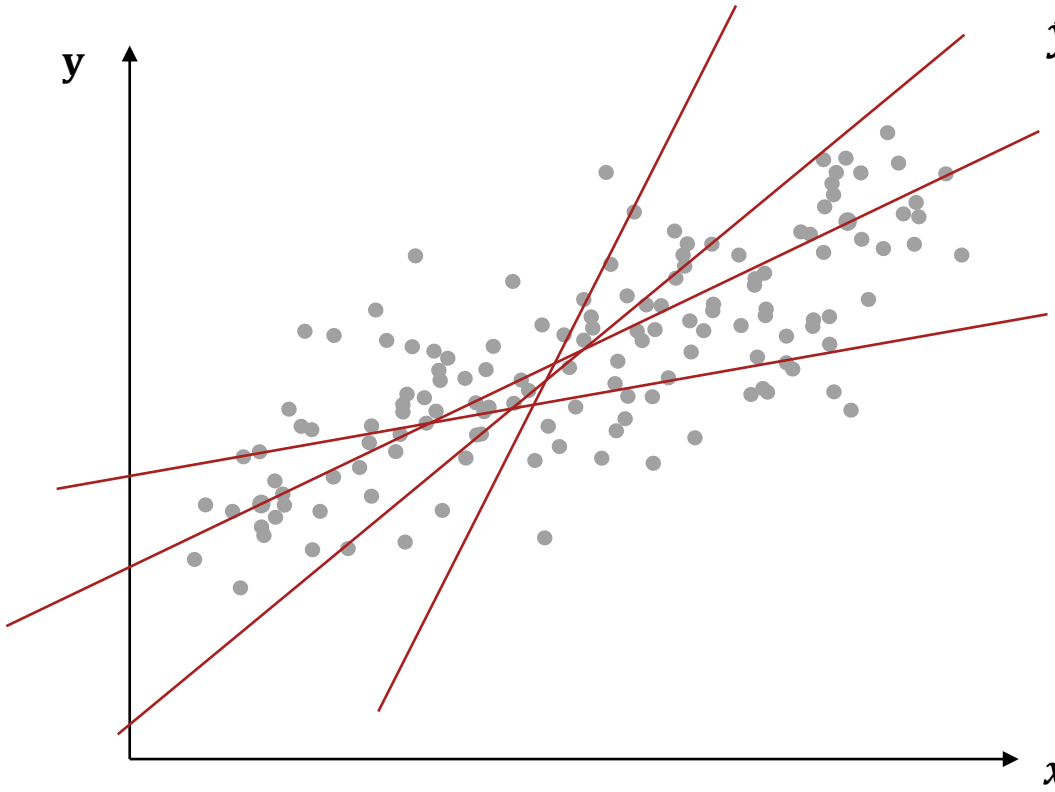
관측된 Data에서 회귀모형으로 설명되는 부분

잔차(Residual)
: 회귀모형으로 설명되지 않는 부분

- 우리가 추정할 식; 우리 Data로 최대한 진리의 관계를 설명하고자 하는 식
- 나의 모델에서 모수인 β를 추정한 기울기(Slope)를 **회귀계수(Coeffieicnt)**라고 하고, 모수와의 표현 상 구분을 위해 햇(Hat) 표시를 함
- True Model에서 설명되지 않는 부분인 오차항(Error term)은 실제 관측할 수 없는 오차이지만, 우리 모델에서 설명되지 않는 부분은 잔차(Residual)라 부르고, 잔차의 크기가 어느 정도인지 측정할 수 있음

회귀분석의 기울기와 절편은 어떻게 찾을까?

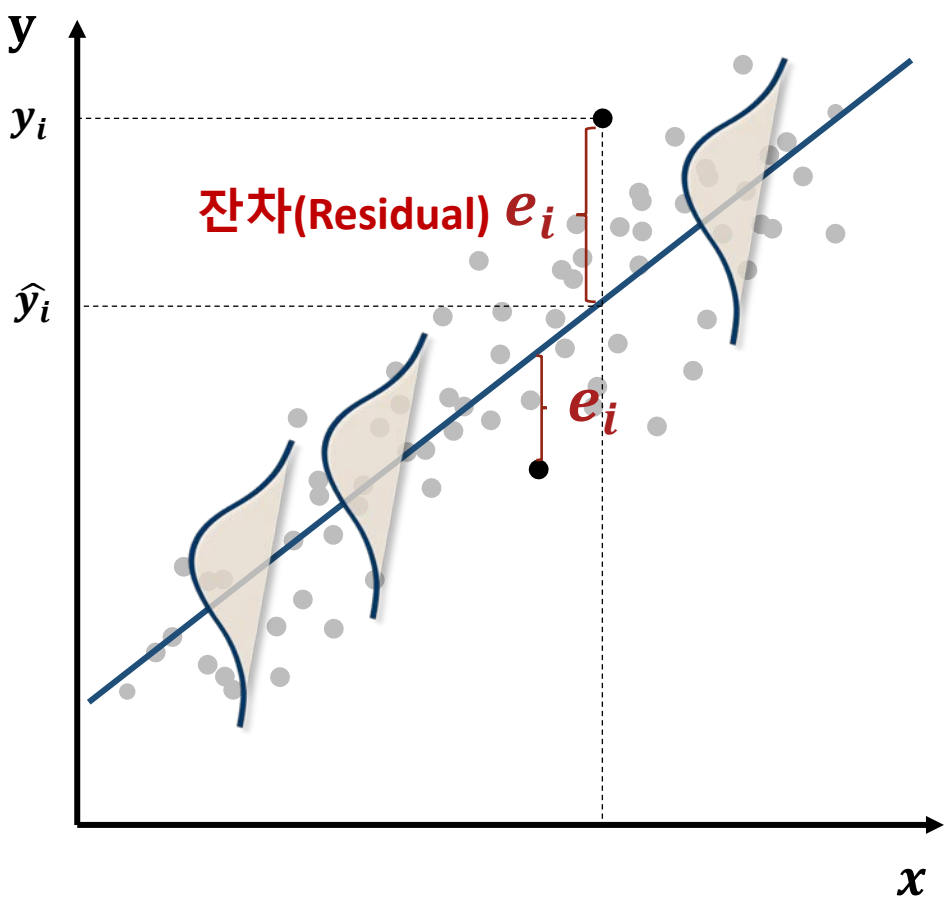
그렇다면, 어떻게 회귀라인을 추정할까? 아래 여러 개의 선형 모형 중 가장 좋은 직선은 무엇일까 ?



$$\hat{y}_i = \widehat{\beta_0} + \widehat{\beta_1} x_i$$

회귀분석의 기울기와 절편은 어떻게 찾을까?

최소제곱법(OLS : Ordinary Least Square Estimation) : 잔차(Residual)를 2차 함수의 형태로 만들고, 이를 최소화하는 점에서 기울기와 절편 즉, 회귀계수(Coefficient)가 결정됨



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

✓ 잔차(Residual)은 전체에서 설명되는 부분을 제한 값으로 표현됨

$$e_i = y_i - \hat{y}_i = y - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

✓ 만약, 모든 잔차들(e_i)을 더하면 0이 됨 => "잔차의 정규성"

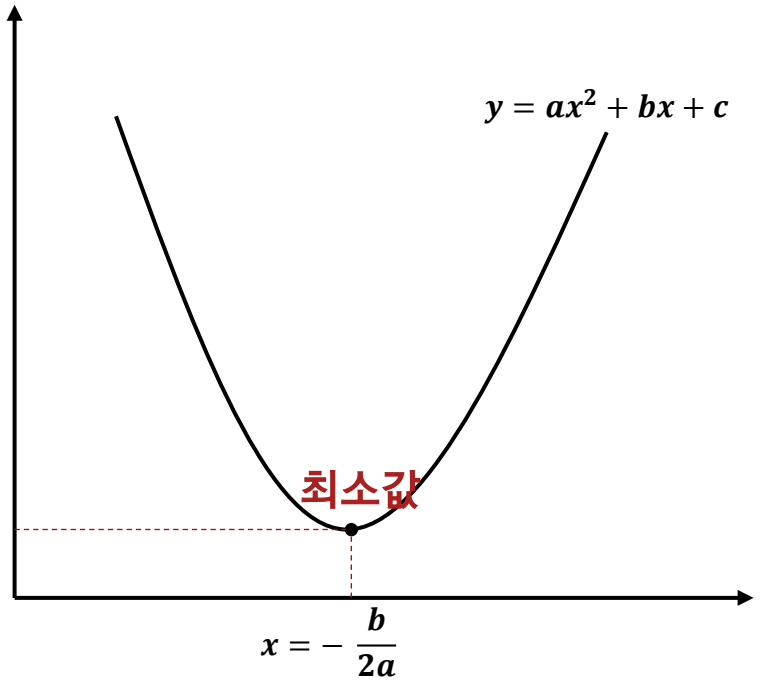
$$\Rightarrow \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} y - \hat{\beta}_0 - \hat{\beta}_1 x_i = 0$$

✓ 그렇기 때문에, 잔차를 제곱해서 Sum하면 2차 방정식 꼴로 나타낼 수 있으므로 최소값을 구할 수 있음

$$\Rightarrow \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} y - \hat{\beta}_0 - \hat{\beta}_1 x_i = 0$$

회귀분석의 기울기와 절편은 어떻게 찾을까?

2차 함수의 극대값(최대값 혹은 최소값)과 미분



$$y = ax^2 + bx + c$$

미분

$$\Rightarrow \frac{\Delta y}{\Delta x} = 2ax + b$$

2차 함수에서 1차 미분한 값을 0으로 만드는 X값이 최적값

$$\Rightarrow 2ax + b$$

$$\Rightarrow 2ax + b = 0$$

접선의 방정식에서 x=0일 때의 기울기

$$\Rightarrow \frac{\Delta y}{\Delta x_{x=0}} = 2a * 0 + b = b$$: y절편 접선의 기울기

만약, Error를 2차함수이면서, 아래로 볼록한 형태로 나타낼 수 있다면, 미분을 통해 회귀분석에서 Error를 최소화하는 기울기(slope)와 절편(intercept)를 구할 수 있음!

회귀분석의 기울기와 절편은 어떻게 찾을까?

회귀계수(Coefficient)

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}, \qquad \widehat{\beta_0} = \bar{y} - \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}\bar{x}$$

증명(Proof)

$y_i = \widehat{\beta_0} + \widehat{\beta_1} x_i + e_i$ → $\widehat{y_i}$

$$y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i = e_i \quad (1)$$

양변을 제곱함

$$(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i)^2 = {e_i}^2 \quad (2)$$

전체 표본의 합을 구함

$$\sum_{i=1}^{n} (y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i)^2 = \sum_{i=1}^{n} {e_i}^2 \quad (3)$$

식 (3)의 우변을 $\widehat{\beta_1}$에 대해 미분하면,

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

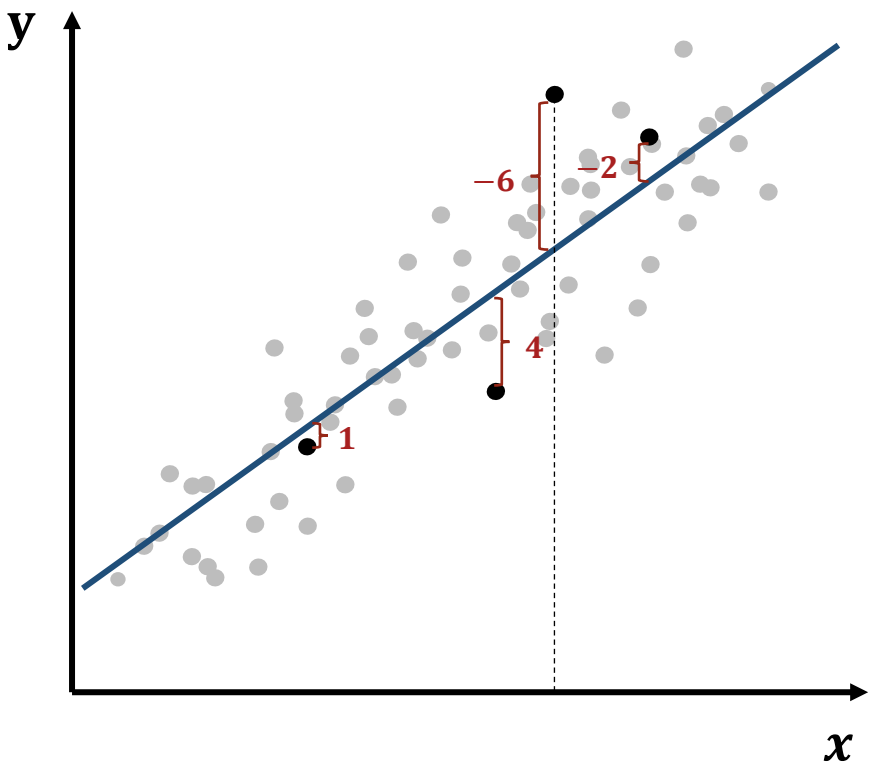
식 (3)의 우변을 $\widehat{\beta_0}$에 대해 미분하면,

$$2\sum_{i=1}^{n} (y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i) = 0 \iff \widehat{\beta_0} = \bar{y} - \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}\bar{x}$$

잔차의 제곱합을 최소로 한다는 조건으로 기울기에 해당하는 $\widehat{\beta_1}$과 절편에 해당하는 $\widehat{\beta_0}$을 구했으므로 회귀선(Regression line)을 그릴 수 있음

참고 – 왜 “절대값”이 아니라 “제곱”을 쓸까?

한때, “최소제곱오차”가 아니라 “최소절대오차”를 쓰면 되지 않느냐 라는 논쟁이 통계학자 사이에서 제기되었으나, 물리학자들이 “최소제곱오차”가 더 성능이 우수하다고 증명한 바 있음



- 최소절대오차 : $|-6| + |-2| + |4| + |1| = 13$

  a) -2에 오차 1만큼 줄일 경우,
  $|-6| + |-1| + |4| + |1| = 12$

  b) -6에 오차 1만큼 줄일 경우,
  $|-5| + |-2| + |4| + |1| = 12$

- 최소제곱오차 : $(-6)^2 + (-2)^2 + (4)^2 + (1)^2 = 57$

  a) -2에 오차 1만큼 줄일 경우,
  $(-6)^2 + (-1)^2 + (4)^2 + (1)^2 = 54$

  b) -6에 오차 1만큼 줄일 경우,
  $(-5)^2 + (-2)^2 + (4)^2 + (1)^2 = 46$

최소절대오차는 멀리 있는 오차와 가까이 있는 오차의 줄어드는 폭이 같으나, 제곱오차는 더 멀리 떨어진 오차를 줄이려고 할 것이므로 가장 잘 설명하는 라인(회귀라인)을 찾는다는 점에서 더 우수한 성능을 나타냄

결론적으로, 이것만 기억하자.

1. 선형회귀모형은 선형 관계 추정이다.

2. 실제 Y(타겟변수)는 모형으로 "설명되는 부분 $\hat{y}_i$"과 "설명하지 못한 부분 잔차(e_i)" 으로 구분된다.

$$y_i = \hat{y}_i + e_i, \quad \hat{y}_i = \widehat{\beta_0} + \widehat{\beta_1} x_i$$

3. 잔차(Residual, e_i)는 정규성을 띄어야 한다. 잔차 정규성을 띄지 않는 변수의 회귀분석 결과는 믿을 수 없다 !

Lecture 7-2

단순회귀분석 예시

단순 회귀분석 예시#1

OECD 주요 국가의 1인당 GDP와 자영업자 비중 간 관계

단순 회귀분석 예시#1

OECD 주요 국가의 1인당 GDP와 자영업자 비중 간 관계

```
Call:
lm(formula = gdp ~ Selfemp, data = selfemp)

Residuals:
   Min     1Q Median     3Q    Max
-35092  -9593    390   8657  47019

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    67588       7153   9.449 2.35e-10 ***
Selfemp        -1682        395  -4.259 0.000197 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17320 on 29 degrees of freedom
Multiple R-squared:  0.3848,    Adjusted R-squared:  0.3636
F-statistic: 18.14 on 1 and 29 DF,  p-value: 0.0001972
```

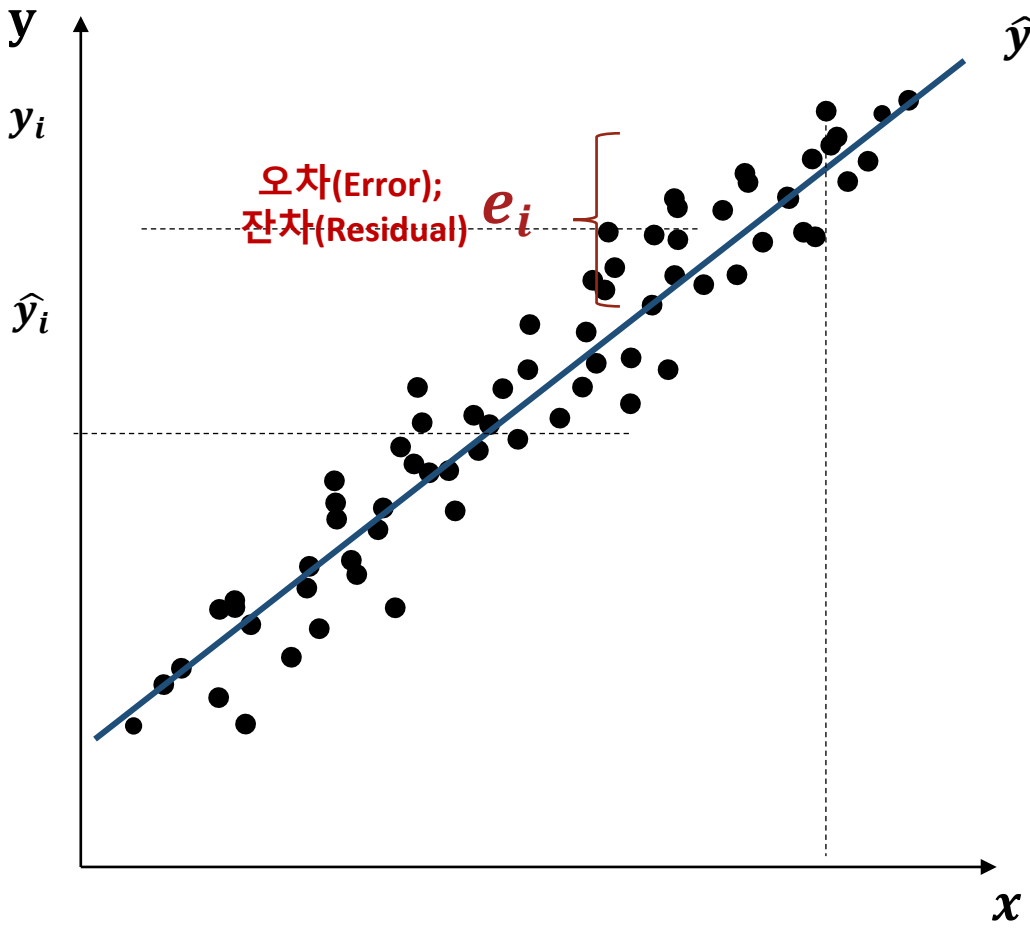
절편, 상수항 : β_0

기울기, 종속변수의 영향 정도 : β_1

모형의 설명력

왜 실제값과 예측값 간 차이가 발생하는 것인가?

선형 회귀모형은 가장 잘 설명할 수 있는 하나의 직선을 긋는 것이므로 각각의 점으로부터 거리가 발생하게 되고, 이 거리가 오차(Error)가 되어 실제값과 예측값 사이에 차이가 발생함



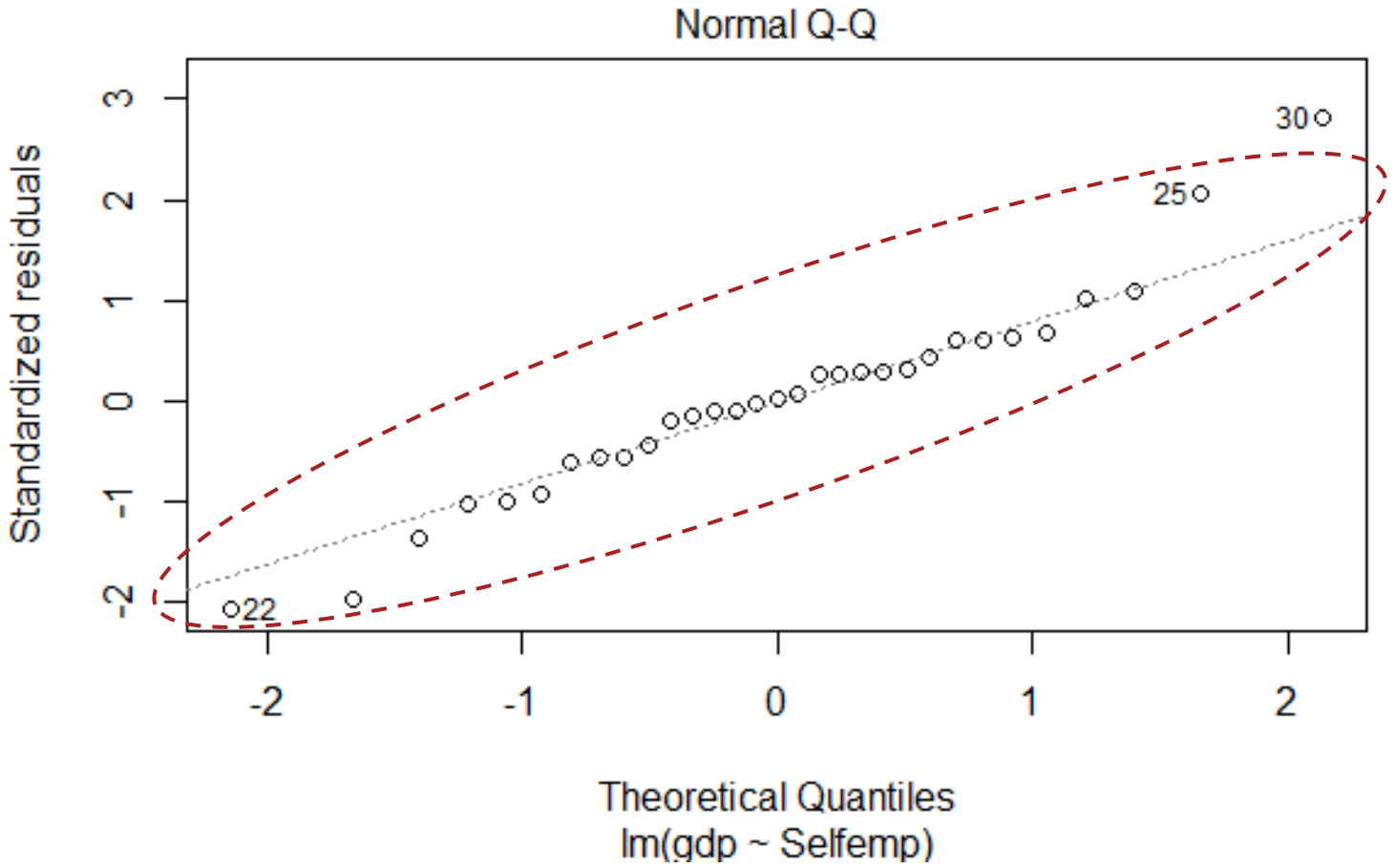
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Best model = The smallest error

각각의 점에서 발생한 오차(Error)들을 평균내면 모형의 평균오차를 구할 수 있고, 여러 대안들이 있다면 평균오차가 가장 적은 모형이 **가장 좋은 모형**이 될 수 있음

잔차(Residual)의 정규성 검증

잔차의 정규성 검증은 명확한 도구가 존재하는 것은 아니며, 일반적으로 Q-Q Plot을 통해 판단함. QQ plot은 가상의 Quantile과 실제 residual 분포를 비교해 45도 선상에 위치하면 정규성을 지닌다고 판단함



Lecture 7-3

다중회귀분석
(Multiple Regression)

다중회귀분석에서 회귀계수의 의미

회귀식

$$\hat{y}_i = 350 + 250x_{1i} - 220x_{2i}$$

- 350: 상수항(Constant)
- 250: x_1의 계수(Coefficient)
- -220: x_2의 계수(Coefficient)

회귀계수 해석

- 상수항(Constant) = 독립변수 x들이 모두 0일 때의 값
- x_1의 계수(Coefficient) = 다른 독립변수들의 효과를 고정시킨 가운데, x_1이 1변할 때, y의 변화량
- x_2의 계수(Coefficient) = 다른 독립변수들의 효과를 고정시킨 가운데, x_2이 1변할 때, y의 변화량

회귀분석 예시#2

도요타(Toyota) 코롤라(Corolla) 중고차 가격 결정 모형

회귀분석 예시#2

도요타(Toyota) 코롤라(Corolla) 중고차 가격 결정 모형

절편

기울기

```
Call:
lm(formula = Price ~ ., data = toyota_train)

Residuals:
     Min       1Q   Median       3Q      Max
-10775.1   -744.9    -27.1    751.5   6362.3

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.049e+04  1.656e+03  -6.334 3.61e-10 ***
Age           -1.156e+02  2.974e+00 -38.859  < 2e-16 ***
KM            -1.678e-02  1.521e-03 -11.032  < 2e-16 ***
FuelTypeDiesel 3.201e+03  6.061e+02   5.281 1.58e-07 ***
FuelTypePetrol 1.833e+03  4.154e+02   4.413 1.13e-05 ***
HP             5.204e+01  6.231e+00   8.351 2.26e-16 ***
MetColor       2.459e+02  8.583e+01   2.865  0.00426 **
Automatic      1.747e+02  1.791e+02   0.975  0.32978
CC            -3.470e+00  6.036e-01  -5.749 1.19e-08 ***
Doors         -1.085e+02  4.652e+01  -2.333  0.01982 *
Weight         2.545e+01  1.551e+00  16.405  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1261 on 996 degrees of freedom
Multiple R-squared:  0.8776,    Adjusted R-squared:  0.8764
F-statistic: 714.4 on 10 and 996 DF,  p-value: < 2.2e-16
```

Lecture 7-4

회귀분석
모형 평가

회귀모형 성능은 어떻게 측정할까?

회귀분석의 잔차(Residual)의 합은 항상 0 이므로 단순 평균을 내면 0이 나옴. 따라서, 모형평가 지표로는 평균절대오차(MAE) 또는 제곱근평균제곱오차(RMSE)가 활용됨

평균절대오차(Mean Absolute Error, MAE)

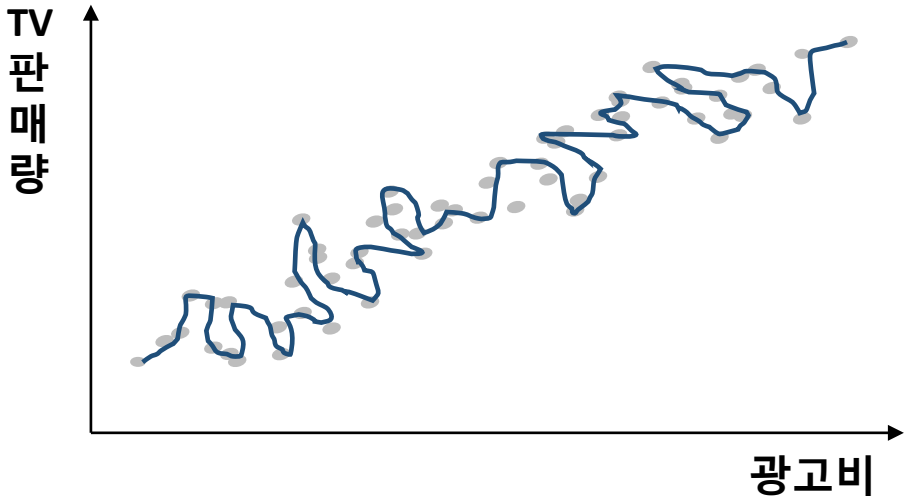
$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\, y_i - \hat{y}_i \,| = \frac{1}{n}\sum_{i=1}^{n}|\, e_i \,|$$

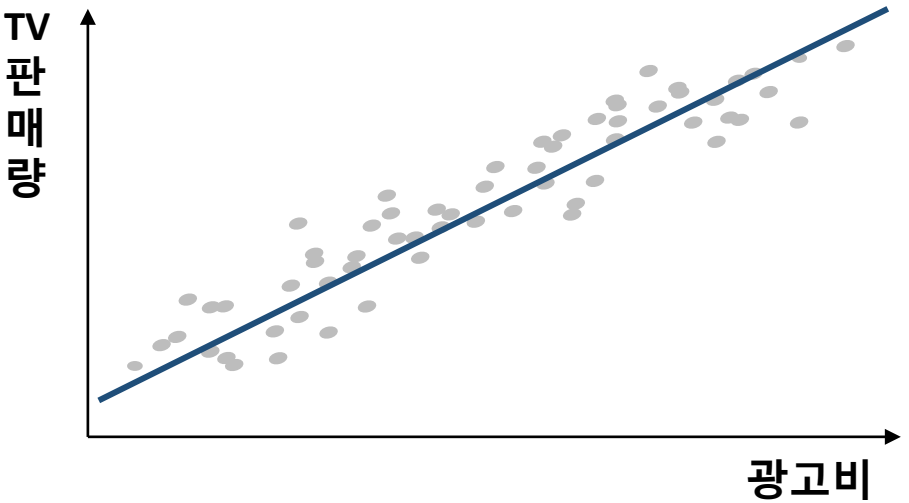
제곱근평균제곱오차(Root Mean square Error, RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}e_i^{\,2}}$$

과적합 문제(Overfiting problem)

왜 그럼 오차(Error)를 허용하면서 직선으로 추정하는 것인가? 모든 점을 지나도록 곡선으로 추정하면 안되는 것인가? 다음의 예시를 보자.





과적합 문제(Overfitting Problem)

- ✓ 모형을 학습시키는 목적은 표본으로부터 일반적인 결과를 도출해 모집단의 특성을 추정하고, 예측하는 것
- ✓ 과적합의 문제가 생기면, 모형의 적합도는 높으나 새로운 자료(Data)에 대한 예측 성능이 좋지 않음
- ✓ 우리 목적은 Training을 잘하는 것도 중요하지만 Prediction을 잘하는 것이 더욱 중요함
- ✓ 따라서, 선형 모형의 목적은 오차를 어느정도 허용하더라도 선형관계를 통해 예측성능을 높이고자 하는 것임

"Linear is beautiful"

Lecture 7-5

가변수의 이해

가변수(Dummy variable) 이해

범주형(Factor, Character) 변수를 모형에 반영할 때는 가변수/더미변수(Dummy variable)를 이용해서 정의함

1) 범주가 2개 짜리 변수 (ex. 성별)

Id	성별_string	성별_numeric
1	남성	0
2	여성	1
3	여성	1
4	남성	0
5	여성	1

$$y_i = \widehat{\beta_0} + \widehat{\beta_1} * 성별_{dummy} + e_i$$

- 남성($성별_{dummy} = 0$) : $y_i = \widehat{\beta_0} + e_i$

- 여성($성별_{dummy} = 1$) : $y_i = \widehat{\beta_0} + \widehat{\beta_1} + e_i$

여성이 남성보다 평균적으로 $\widehat{\beta_1}$ 만큼 y_i에 더 영향을 미친다!

※ 더미 변수에서 0이 되는 범주를 Reference category 라고 함. 여기서는 "남성"이 Reference 임.

가변수(Dummy variable) 이해

범주형(Factor, Character) 변수를 모형에 반영할 때는 가변수/더미변수(Dummy variable)를 이용해서 정의함

2) 범주가 3개 짜리 변수 (ex. 연료타입)

Id	연료_string	연료_경유	연료_휘발유
1	CNG	0	0
2	경유	1	0
3	경유	1	0
4	휘발유	0	1
5	CNG	0	0

※ 여기서 Reference는 ( ) 임.

$$y_i = \widehat{\beta_0} + \widehat{\beta_1} * 연료_{Diesel} + \widehat{\beta_2} * 연료_{Petrol} + e_i$$

- CNG(0, 0) : $y_i = \widehat{\beta_0} + e_i$
- 경유(Diesel)(1, 0) : $y_i = \widehat{\beta_0} + \widehat{\beta_1} + e_i$
- 휘발유(Petrol)(0, 1) : $y_i = \widehat{\beta_0} + \widehat{\beta_2} + e_i$

CNG에 비해 경유는 y_i에 $\widehat{\beta_1}$ 만큼 더 큰 영향을 미친다.
CNG에 비해 휘발유는 y_i에 $\widehat{\beta_2}$ 만큼 더 큰 영향을 미친다.

가변수(Dummy variable) 이해

범주형(Factor, Character) 변수를 모형에 반영할 때는 가변수/더미변수(Dummy variable)를 이용해서 정의함

- **범주가 5개 짜리 변수는 몇 개의 더미 변수가 생성될까?** **( )**

- **위에서 생성된 더미 변수의 Reference는 어떤 값을 가질까?** **( )**

Lecture 7-6

모형 성능의 개선

모형개선(Model Improvement)

자료 추가

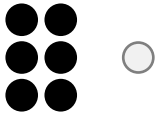
Adding more data

교차 검증법

Cross Validation

피처 엔지니어링

Feature Engineering



이상치 제거

Removing Outlier

상호 작용항

Interaction Term

다중공선성

Multi-collinearity

모형개선(Model Improvement)

자료 추가

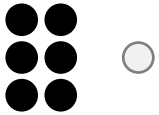
Adding more data

교차 검증법

Cross Validation

피처 엔지니어링

Feature Engineering



이상치 제거

Removing Outlier

상호 작용항

Interaction Term

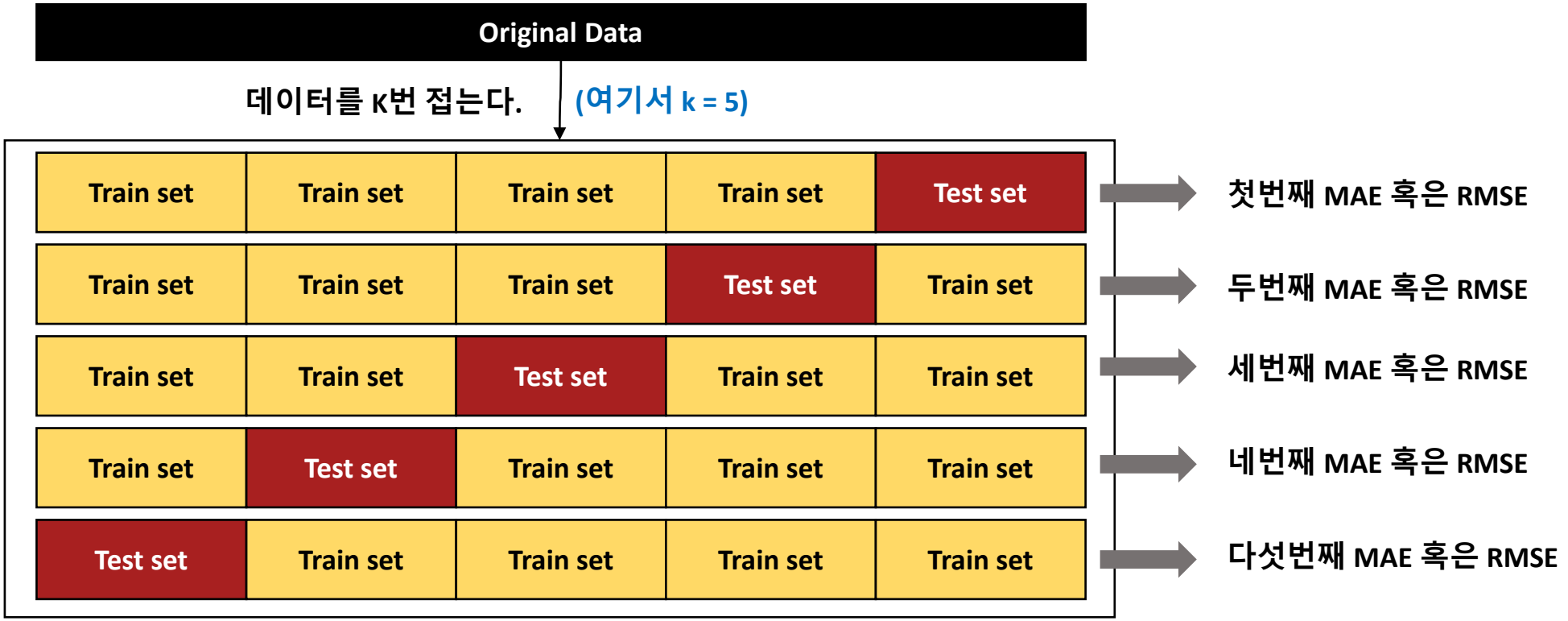
다중공선성

Multi-collinearity

"추가적인 시간 및 비용이 소요될 수 있음"

모형개선#1 – K-겹 교차 검증법(Cross Validation)

모형 개선을 위해 우리는 수 차례 모형을 개선시키는 데, 이 과정에서 고정된 Train set과 Test set을 활용한다면 결국 과적합(Overfitting) 문제가 발생할 수 있음.



모형의 총 MAE = mean(첫번째 MAE, 두번째 MAE, …, 다섯번째 MAE)

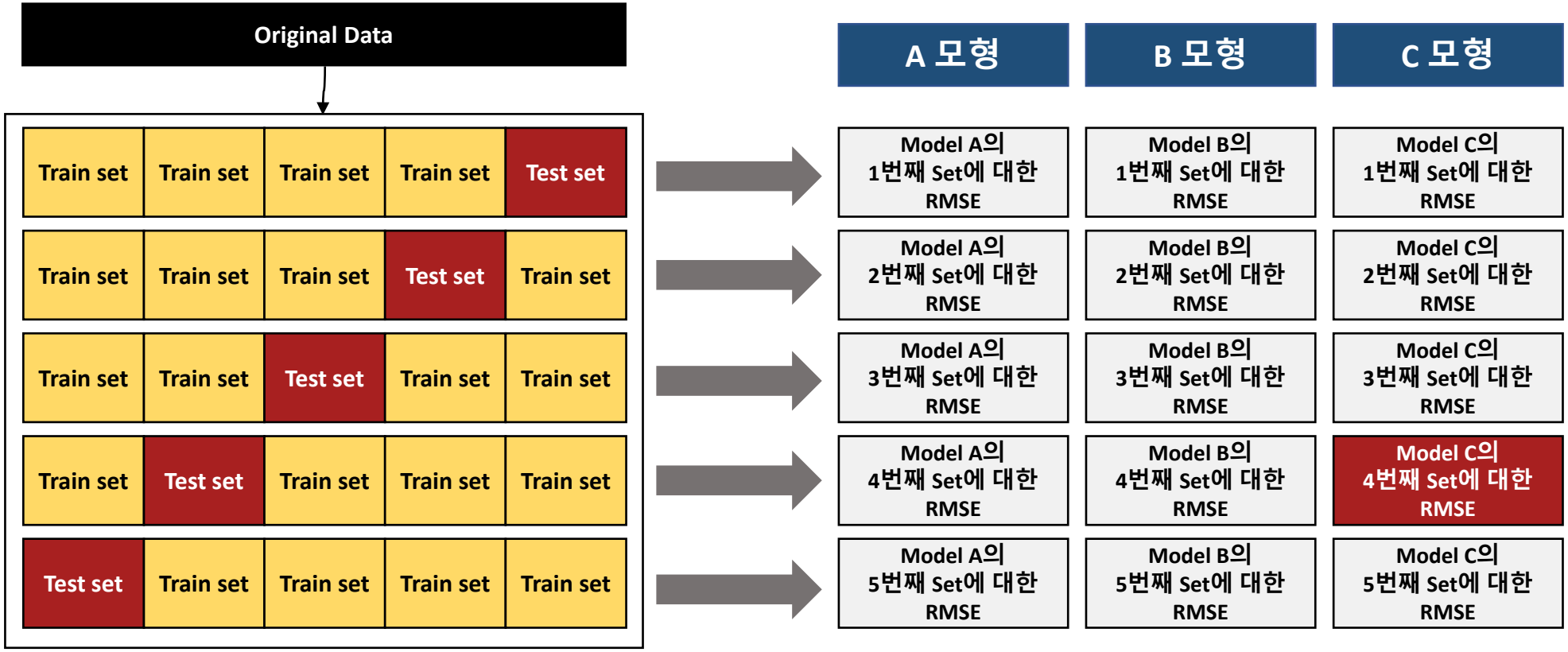
모형의 총 RMSE = mean(첫번째 RMSE, 두번째 RMSE, …, 다섯번째 RMSE)

모형개선#1 – K-겹 교차 검증법(Cross Validation)

1) 데이터를 Random하게 k개의 같은 크기로 쪼갬. 그럼 k개의 Folded set이 나옴
2) 첫번째 folded set에서 K-1개의 데이터를 Training set으로 이용하여 모형을 학습시킴
3) 나머지 1개의 데이터를 Test set으로 하여 Y값을 예측(Prediction)
4) 2)~3)번 과정을 K번 반복해 모든 Y값에 대한 예측값(Predicted value)을 찾아냄
5) 1~4번까지 과정을 각각의 후보모형마다 실행함
6) 1-5까지 과정을 반복함
7) 각 단계마다 MAE 혹은 RMSE를 계산함.
8) MAE 및 RMSE를 모아서 가장 작은 값을 나타내는 모형을 선택함

모형개선#1 – K-겹 교차 검증법(Cross Validation)

예) 현재 고려 중인 모형은 **A, B, C 모형 3개**이고, **5-folded 교차검증**을 하려고 한다.
평가 기준은 **RMSE**를 기준으로 최적 모형을 도출하고자 한다.



총 "15개의 모형"을 비교하는 것과 동일함!

모형개선#2 – 피처엔지니어링 (Feature Engineering)

피처 엔지니어링은 주어진 피처(Feature)들을 이용해

해당 도메인에 대한 지식 및 특성 등을 미리 알거나

탐색적 분석을 통해 알게 된 사실을 바탕으로

유의미한 변수를 생성, 선택 및 변환하는 과정

모형개선#2 – 피처엔지니어링 (Feature Engineering)

경험에 의해 이미 알고 있는 사전지식을 기반으로 존재할 수 없는 값에 대해 직접 변환을 취해주는 작업 역시 피처 엔지니어링의 일환이라 볼 수 있음

<예시>

거래처	A 제품 매출액	B 제품 매출액	C 제품 매출액	…
이마트	100,000	60,000	230,000	…
롯데마트	200,000	-30,000	480,000	…
홈플러스	150,000	40,000	-110,000	…
⋮	⋮	⋮	⋮	⋮

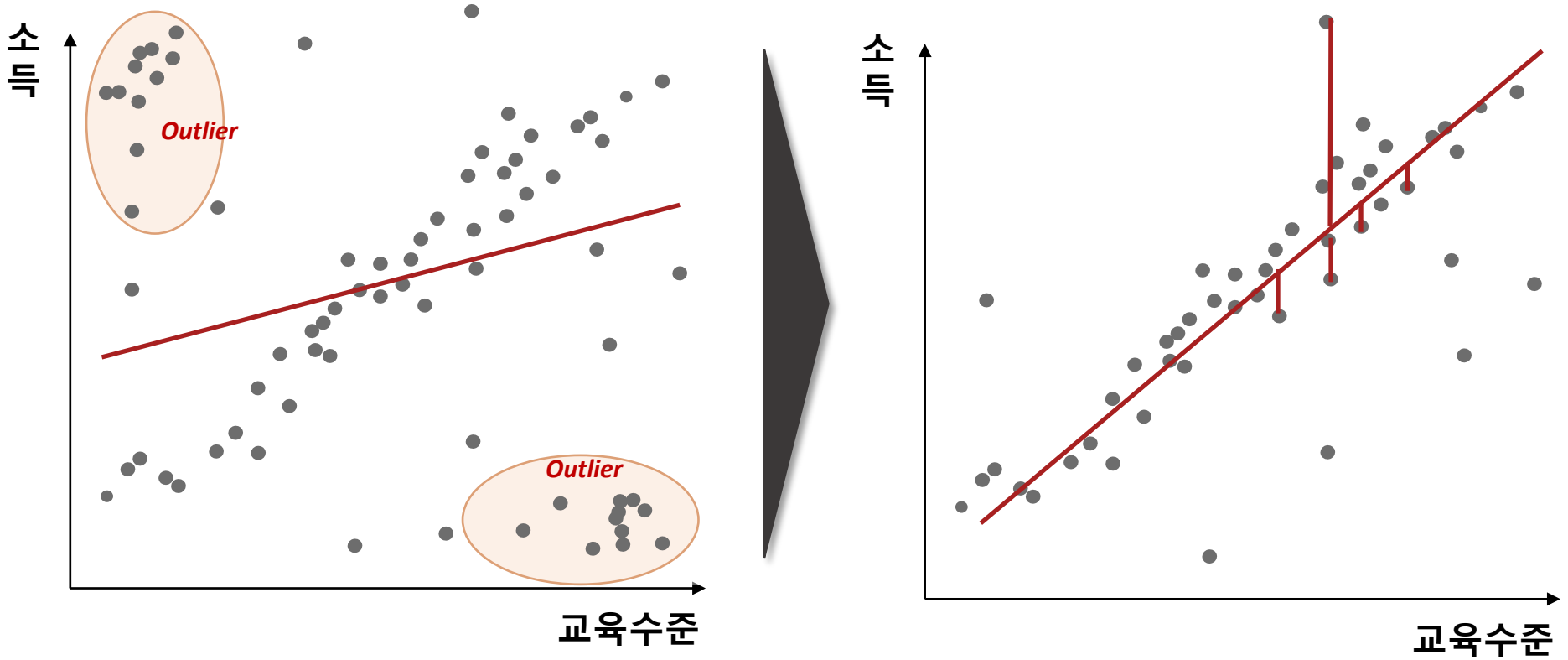
환입으로 인해 변수의 값이 음수(-)인 값이 존재함

거래처	A 제품 매출액	B 제품 매출액	C 제품 매출액	…
이마트	100,000	60,000	230,000	…
롯데마트	200,000	0	480,000	…
홈플러스	150,000	40,000	0	…
⋮	⋮	⋮	⋮	⋮

사전 지식을 바탕으로, 매출액이 음수(-)인 값은 모두 0으로 바꿈

모형개선#3 – 이상치 발견 (Outlier detection)

이상치(Outlier)는 데이터 상에 존재는 하지만 일반적이지 않은 분포나 형태를 나타내고 있는 자료를 의미하며, 자료의 수가 많으면 이상치의 영향이 줄어드나 이상치가 과도하게 많으면 모형성능이 떨어짐



모형개선#3 – 이상치 발견 (Outlier detection)

이상치 (Outlier)

제거

이상치를 제거하면 모형 성능을 개선할 수도 있지만 이상치가 존재하는 관측치 자체가 모형에 투입되지 못하므로 정보의 손실 발생

모형에 투입되지 못함

ID	소득	나이	…
1	350	42	…
2	400	*339 (outlier)*	…
⋮	⋮	⋮	⋮

조정

이상치를 조정하는 방법은 여러가지가 있으며, 대표적으로 해당 변수의 분포 중 양측 90%에 해당하는 데이터로 이상치를 끌어오는 방법이 있음

모형에 투입가능

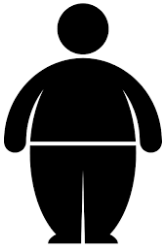
ID	소득	나이	…
1	350	42	…
2	400	*45*	…
⋮	⋮	⋮	⋮

모형개선#4 – 상호작용항 (Interaction Term)

상호작용효과는 서로 독립적인 변수이지만 독립적인 두 변수가 각각 미치는 영향 외에도 두 변수가 동시에 발생했을 때, 시너지(Synergy) 효과가 발생하 는 경우를 말하며 모형에 이를 상호작용항으로 반영함

흡연자

"흡연자이면서 비만 환자라면?"



비만

보험료(Insurance expense)

모형개선#4 – 상호작용항 (Interaction Term)

즉, 두개 이상의 변수가 각각이 주는 효과에다가 두 변수가 결합되면서 미치는 "+@ 효과"를 상호작용효과라 부르며, 이를 회귀모형에 반영한 항이 상호작용항임

수능 점수 = 수학 + 영어 + 수학*영어

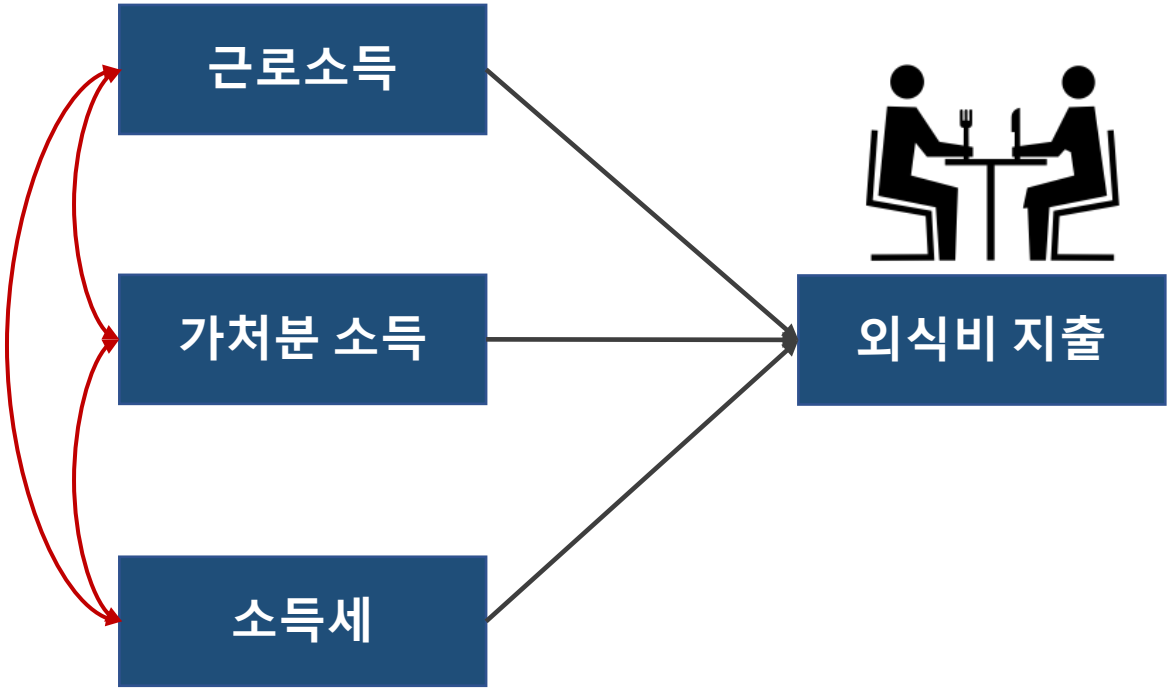
투약 효과 = 약물 A + 약물 B + 약물 A * 약물 B

상호작용항(Interaction term)

모형개선#5 – 다중공선성 (Multicollinearity) 제거

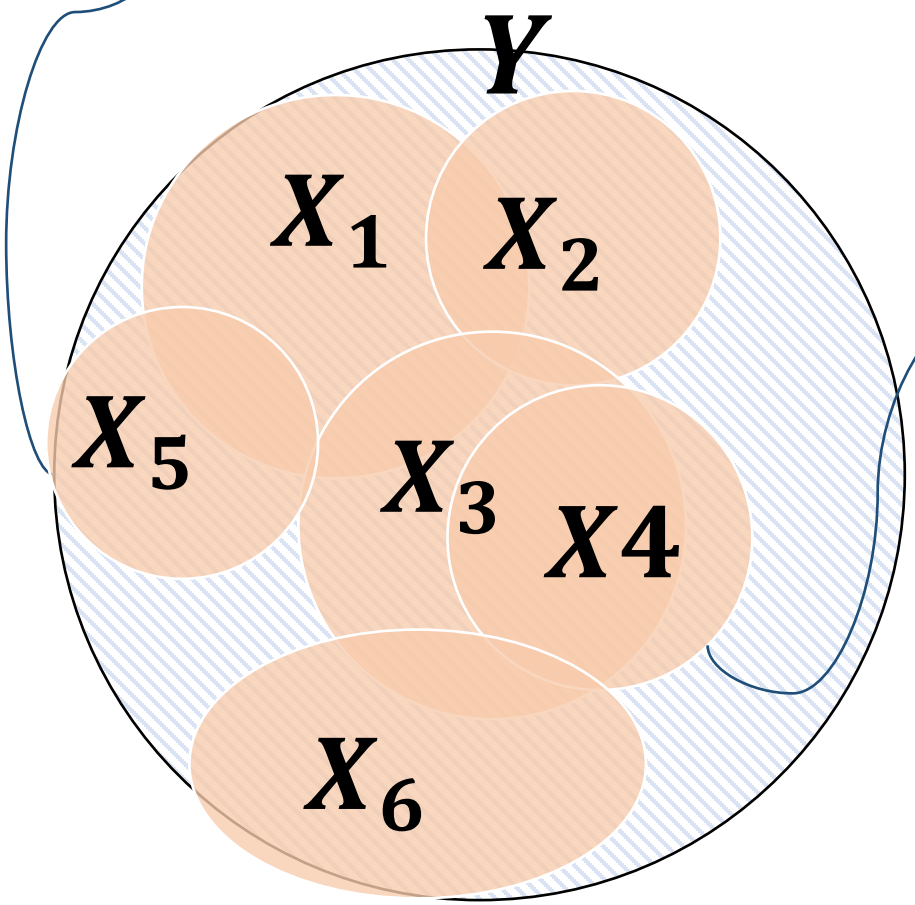
다중공선성은 설명변수로 들어가는 두 변수 간 상관관계가 매우 높은 경우를 말하며, 이 경우 다중공선성을 제거해야 독립변수의 진정한 영향정도를 파악할 수 있고 모형 성능을 높일 수 있음

세 변수 모두 상관관계가 매우 높은데, 설명변수로 모두 써야 하는가?



다중공선성(Multicolinearity)

Y의 변동성 (SST : Total Sum of Squares)



X 변수 간 겹치는 부분이 많다는 의미는 각각의 X가Y를 독립적으로 설명해야 되는데, X 변수 간 (변수 간 상관관계)가 매우 높다는 의미가 됨

➢ **다중공선성(Multicolinearity)**

- ✓ X_3과 X_4는 서로 겹치는 부분이 매우 많고, 사실상 각각이 부분적으로 Y의 변동성을 설명하는 부분은 현저히 적음
- ✓ 이런 경우, "다중공선성"이 존재한다고 볼 수 있음
- ✓ 다중공선성이 존재할 경우, X4가 Y에 미치는 영향이 (+) 임에도 불구하고 부호가 바뀌는 문제 발생할 수 있음

다중공선성(Multicolinearity)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    67588       7153   9.449 2.35e-10 ***
Selfemp        -1682        395  -4.259 0.000197 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17320 on 29 degrees of freedom
Multiple R-squared:  0.3848,    Adjusted R-squared:  0.3636
F-statistic: 18.14 on 1 and 29 DF,  p-value: 0.0001972
```

- 자영업자 비율(Selfemp)을 독립변수로 한 모형에서는 높은 t-value와 p-value를 나타내며 음(-)의 영향을 지님

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  73309.8     8579.0   8.545 2.06e-09 ***
Selfemp_re   -1807.7      435.3  -4.153 0.000264 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17480 on 29 degrees of freedom
Multiple R-squared:  0.3729,    Adjusted R-squared:  0.3513
F-statistic: 17.25 on 1 and 29 DF,  p-value: 0.0002636
```

- 자영업자 비율을 일부 조작한 변수(Selfemp_re)를 독립변수로 한 모형 역시 높은 t-value와 p-value를 나타내며 음(-)의 영향을 지님

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    47529      23982   1.982   0.0574 .
Selfemp_re      5844       6666   0.877   0.3881
Selfemp        -7024       6107  -1.150   0.2598
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17390 on 28 degrees of freedom
Multiple R-squared:  0.4012,    Adjusted R-squared:  0.3585
F-statistic: 9.382 on 2 and 28 DF,  p-value: 0.0007612
```

- 위 두 변수를 같이 넣어 다중회귀모형을 추정하면, 한 변수는 양(+), 한 변수는 음(-)이 되어 단일 회귀분석 결과와 부호가 달라지며, 두 변수 모두 유의하지 않게 됨
- 실제 두 변수의 상관계수는 **0.997**

다중공선성, 어떻게 제거할 수 있을까?

중요한 변수만 선택하는 방법

- 중요하지 않은 변수이 모형에 투입될 필요가 없을 경우 중요하지 않은 변수를 제거하고 모형 Fitting

- 직접 제거
- Stepwise 변수선택법
- LASSO

중요하지 않은 변수의 가중치를 낮추는 방법

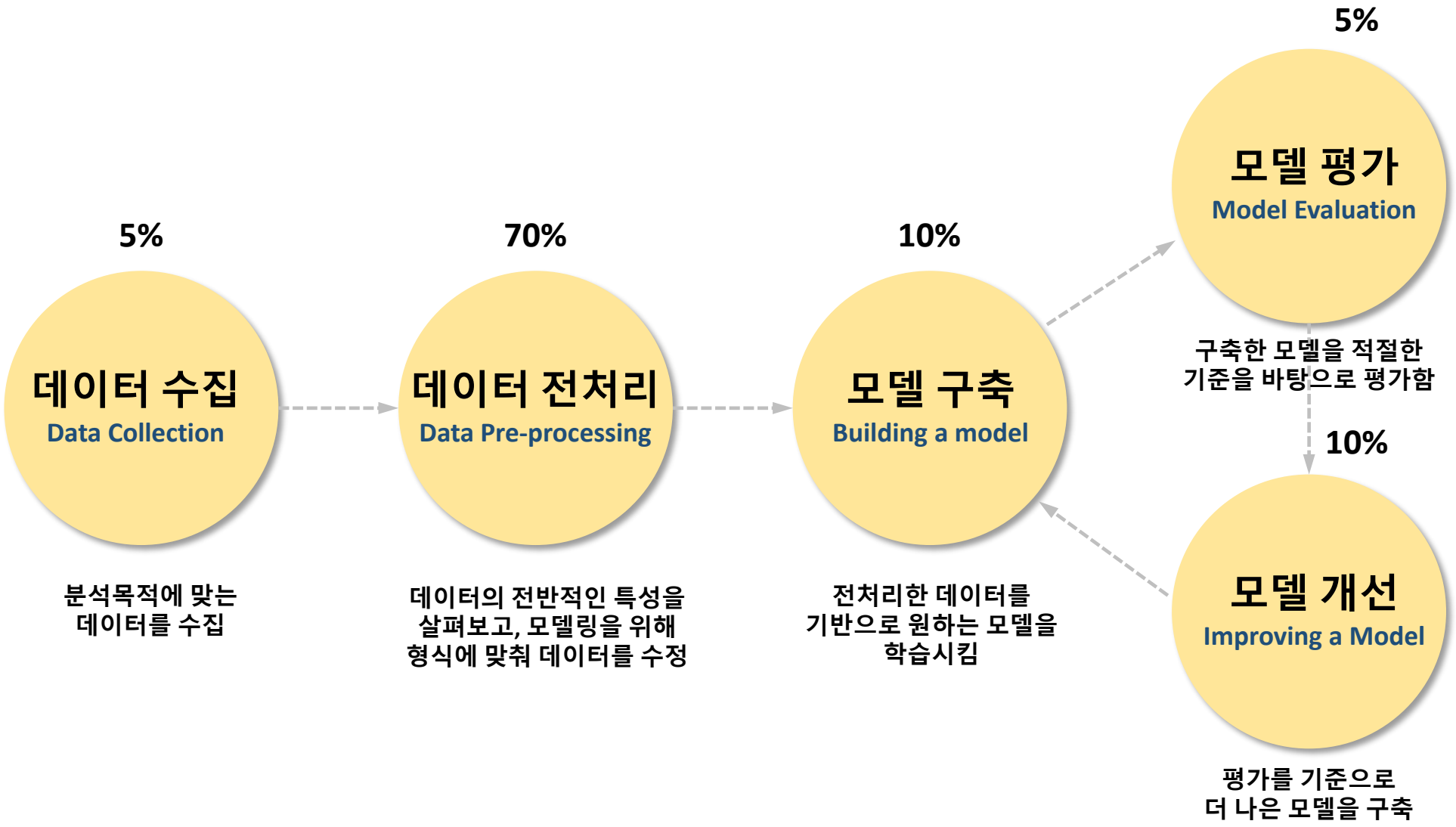
- 중요하지는 않지만 모형에 남아있을 필요가 있을 경우, 중요하지 않은 변수의 가중치를 줄이거나 고차원 변수들을 저차원으로 통합하여 추정하는 방법

- PCA
- Ridge
- ElasticNet

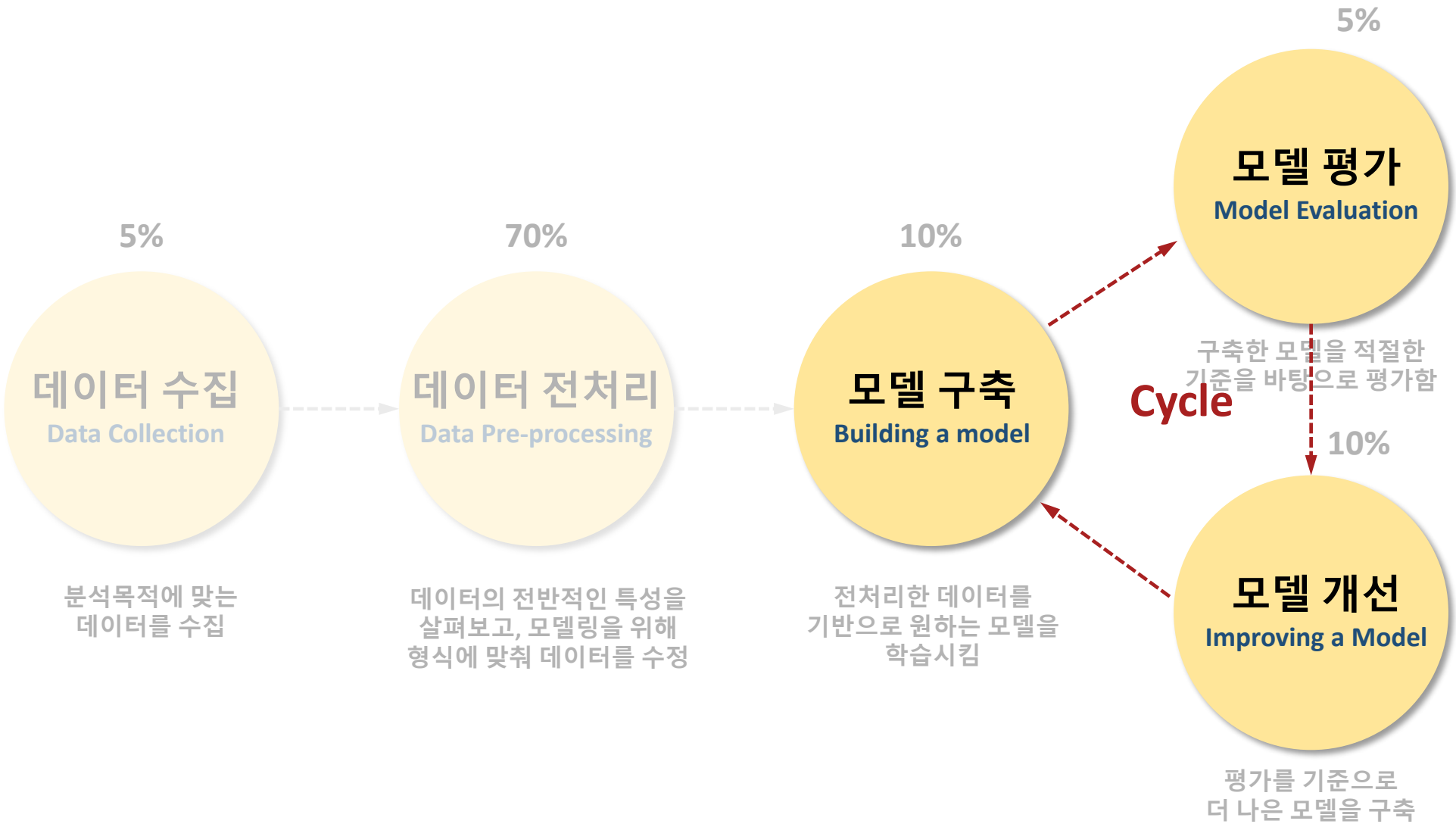
Lecture 7-7

모형 학습 Process

Learning Process

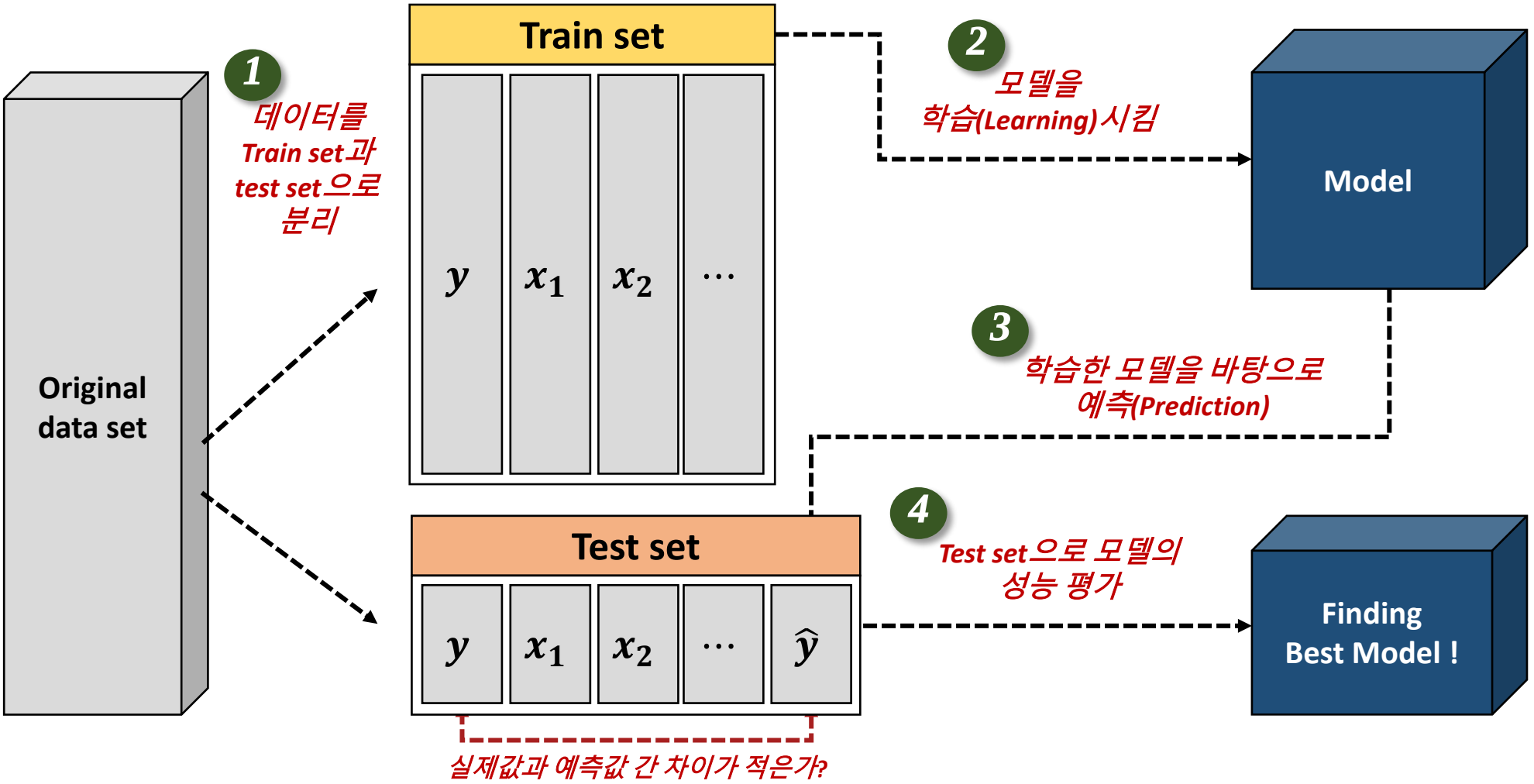


Learning Process



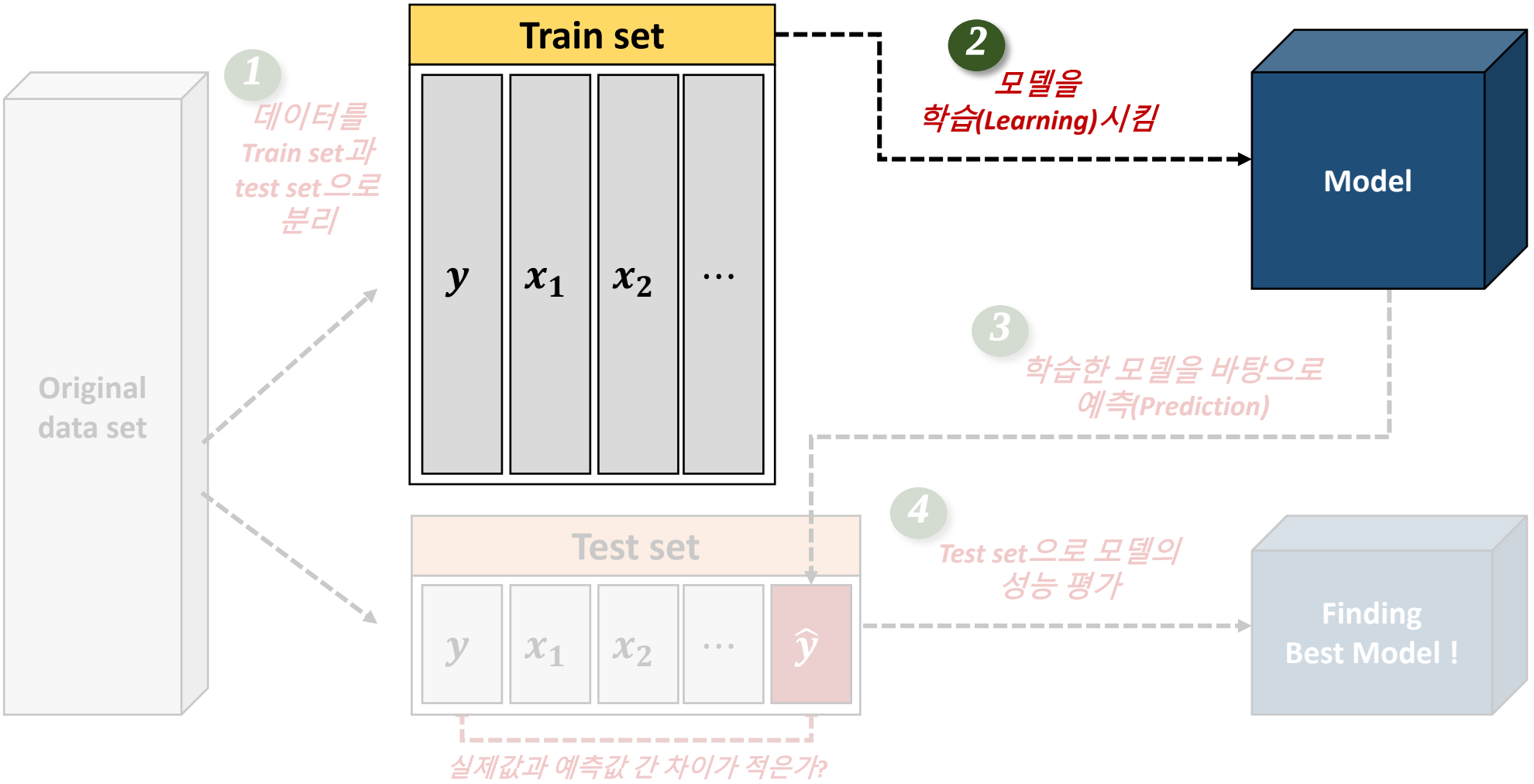
훈련(Training) and 검증(Testing)

예측모형을 도출하기 위해서는 내가 갖고 있는 원래의 자료(Original data set)를 훈련데이터(train set)와 검증데이터(test set)으로 나누고, 이를 지속적으로 학습시켜 최적의 모형을 찾아야 함



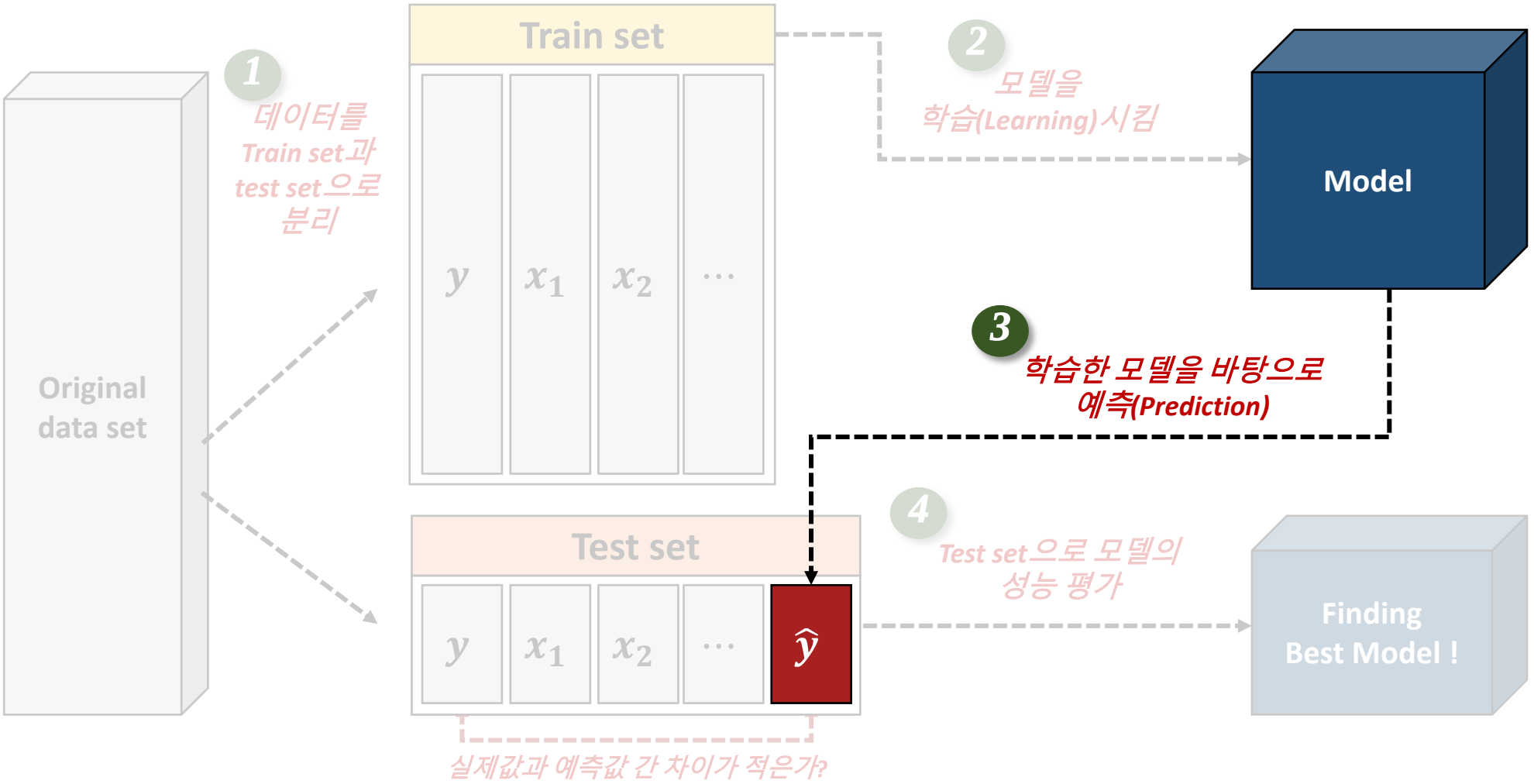
훈련(Training) and 검증(Testing)

예측모형을 도출하기 위해서는 내가 갖고 있는 원래의 자료(Original data set)를 훈련데이터(train set)와 검증데이터(test set)으로 나누고, 이를 지속적으로 학습시켜 최적의 모형을 찾아야 함



훈련(Training) and 검증(Testing)

예측모형을 도출하기 위해서는 내가 갖고 있는 원래의 자료(Original data set)를 훈련데이터(train set)와 검증데이터(test set)으로 나누고, 이를 지속적으로 학습시켜 최적의 모형을 찾아야 함



훈련(Training) and 검증(Testing)

예측모형을 도출하기 위해서는 내가 갖고 있는 원래의 자료(Original data set)를 훈련데이터(train set)와 검증데이터(test set)으로 나누고, 이를 지속적으로 학습시켜 최적의 모형을 찾아야 함