



# 9주차: 모델링과 예측 : 선형회귀

**ChulSoo Park**

School of Computer Engineering & Information Technology  
Korea National University of Transportation



# 학습목표 (9주차)

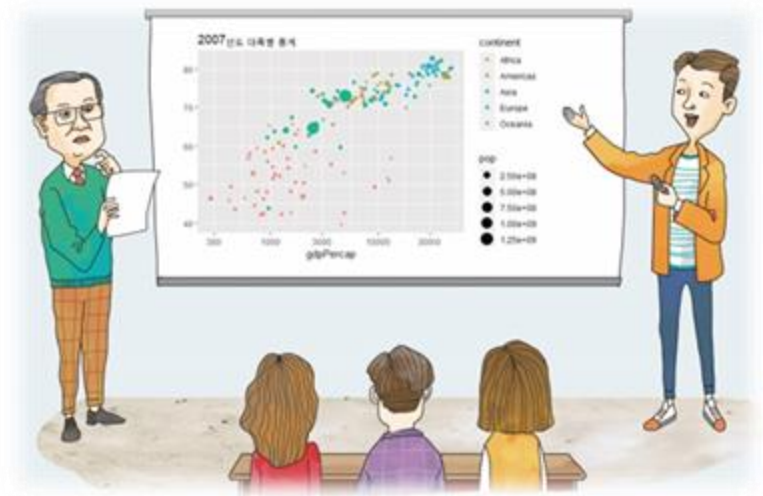
- ❖ 모델링과 예측 이해
- ❖ 회귀 분석 개념 이해
- ❖ 단순회귀분석, 분산 분석(ANOVA) 실행
- ❖ 모델의 통계량 이해
- ❖ t-검정과 분산분석



# 07

## CHAPTER

# 모델링과 예측

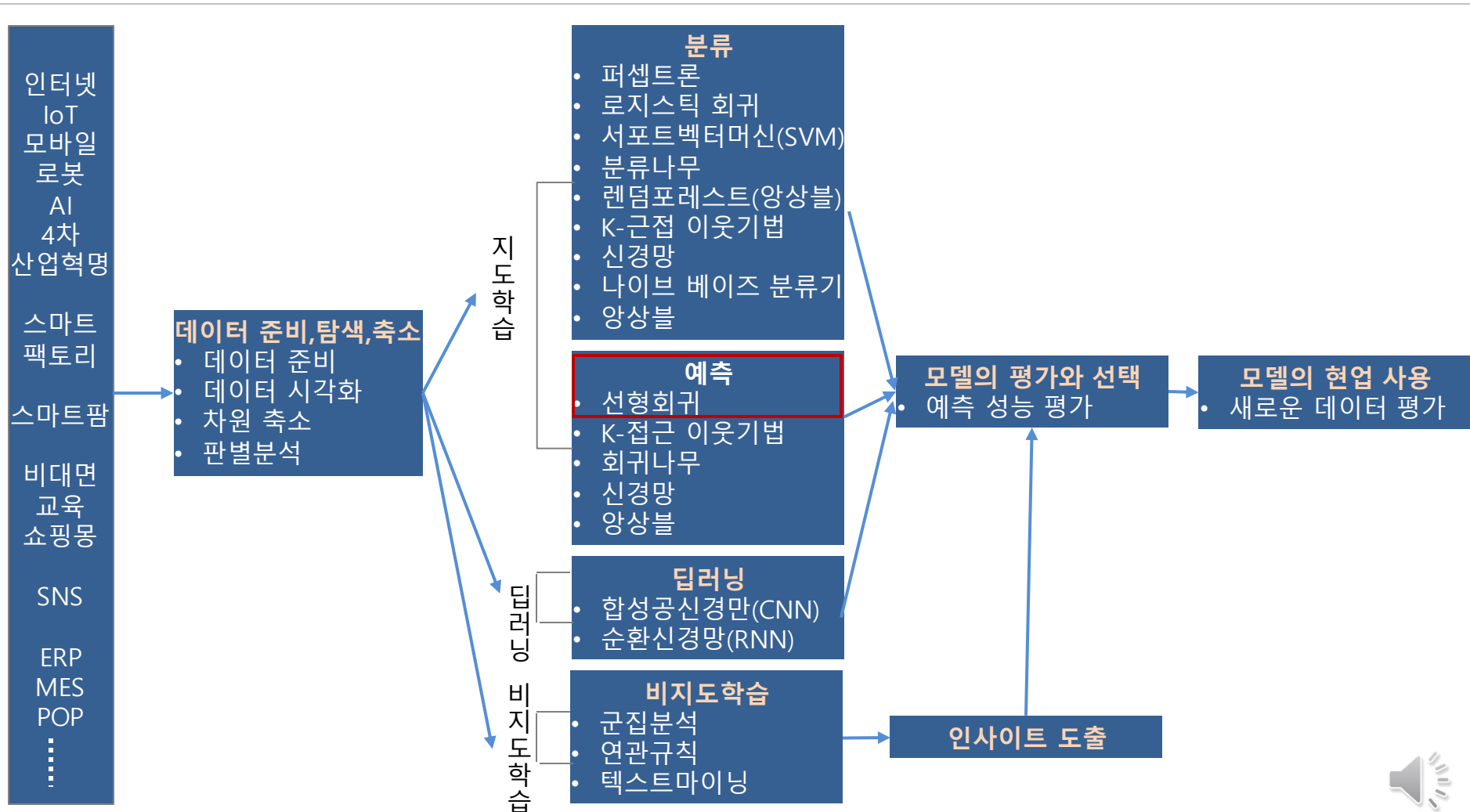


## CONTENTS

- 7.1 모델링과 예측이란?
- 7.2 현실 세계의 모델링
- 7.3 단순 선형 회귀
- 7.4 단순 선형 회귀 적용:cars 데이터
- 7.5 모델의 통계량 해석
- 7.6 다중 선형 회귀
- 7.7 다중 선형 회귀의 적용:trees 데이터
  - t-검정과 분산 분석
  - 요약



## ■ 데이터 분석 Process에서 이번주 교육 위치



## ■ gapminder 데이터의 시각적 탐구(1)

- 세계 여러 나라의 경제 및 복지 수준, 국가나 지역에 따라 나타나는 공통점과 차이점, 아시아 일부 국가들이 지난 수십 년 동안 이룬 눈에 띄는 경제 성장 등 확인할 수 있음
- 또한 country, continent, year, lifeExp, pop, gdpPercap 등 속성들 간의 상호 관계를 분석함으로써 이러한 차이와 변화가 나타난 원인에 대해서 생각해볼 수 있음

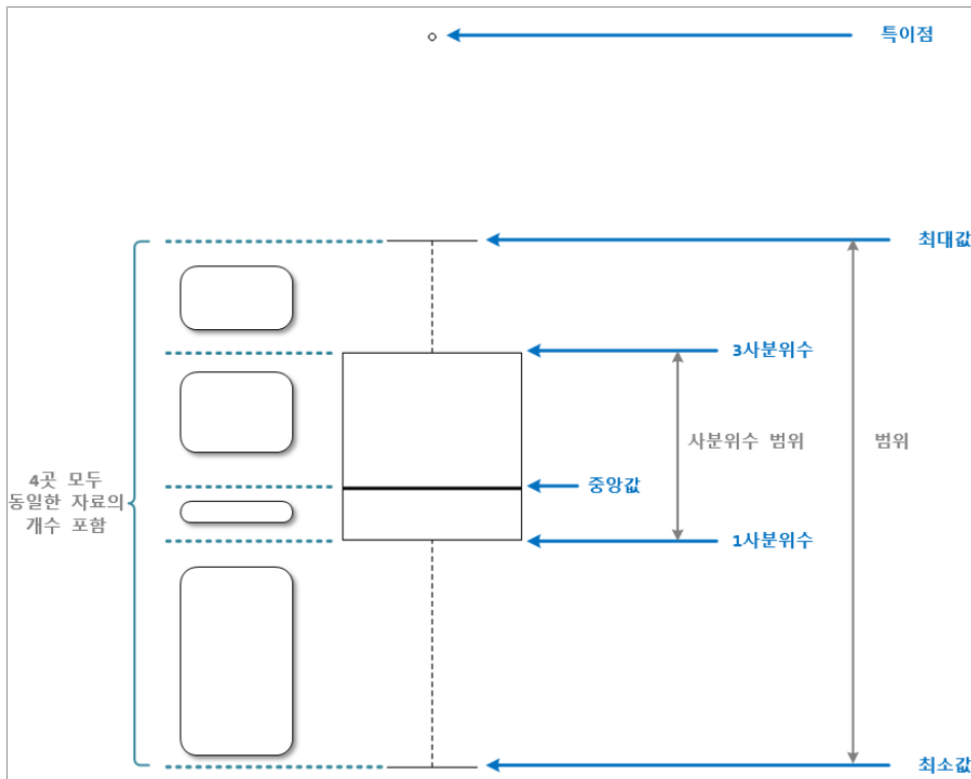
```
> str(gapminder)
tibble [1,704 x 6] (S3: tbl_df/tbl/data.frame)
 $ country   : Factor w/ 142 levels "Afghanistan",...: 1 1 1 1 1 1 1 1 1 1
 1 ...
 $ continent : Factor w/ 5 levels "Africa","Americas",...: 3 3 3 3 3 3 3 3
 3 3 3 ...
 $ year      : int [1:1704] 1952 1957 1962 1967 1972 1977 1982 1987 199
 2 1997 ...
 $ lifeExp   : num [1:1704] 28.8 30.3 32 34 36.1 ...
 $ pop       : int [1:1704] 8425333 9240934 10267083 11537966 13079460
 14880372 12881816 13867957 16317921 22227415 ...
 $ gdpPercap: num [1:1704] 779 821 853 836 740 ...
> head(gapminder)
# A tibble: 6 x 6
  country    continent  year lifeExp    pop gdpPercap
  <fct>      <fct>      <int> <dbl>    <int> <dbl>
1 Afghanistan Asia      1952   28.8  8425333    779.
2 Afghanistan Asia      1957   30.3  9240934    821.
3 Afghanistan Asia      1962   32.0 10267083    853.
4 Afghanistan Asia      1967   34.0 11537966    836.
5 Afghanistan Asia      1972   36.1 13079460    740.
6 Afghanistan Asia      1977   38.4 14880372    786.
```



## ■ geom\_boxplot 함수

- 여러 항목의 분포를 한꺼번에 관찰하는 함수이며, 이상값을 파악하는 데 유용

## ■ 박스 플롯(box plot)의 해석

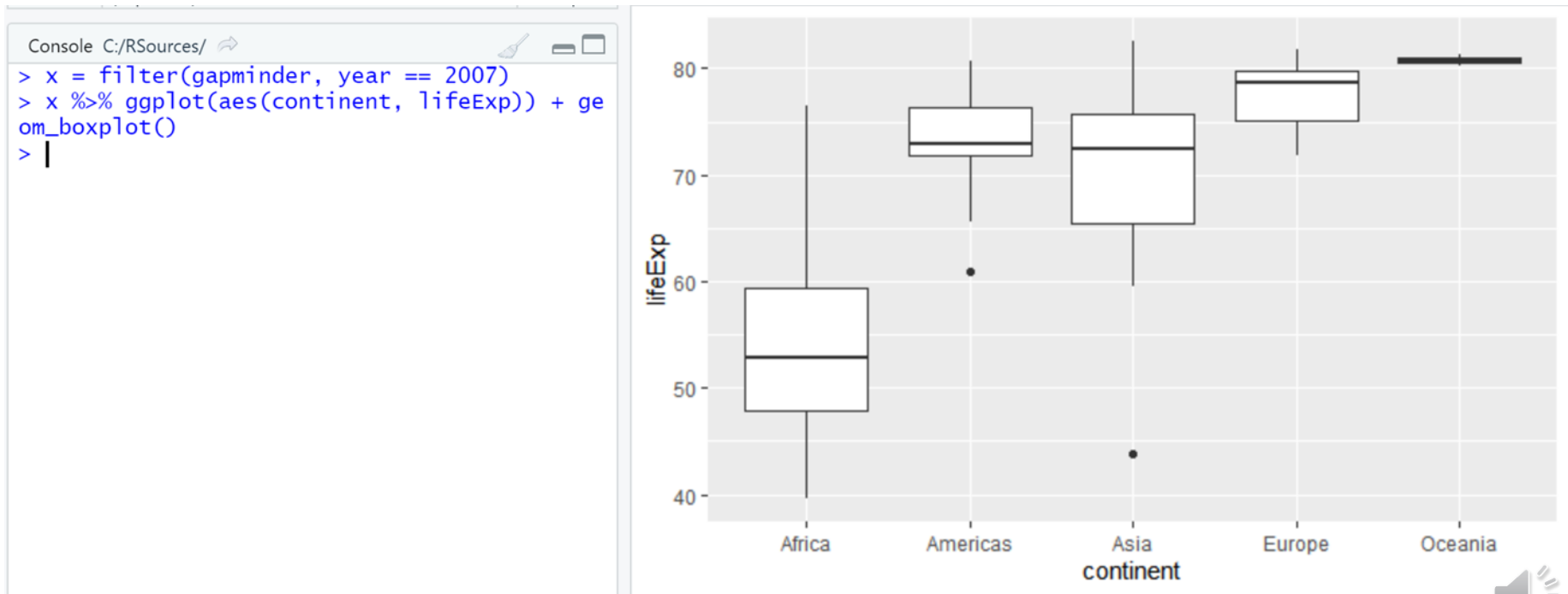


박스 플롯은 박스와 박스 바깥의 선(whisker)으로 이루어져 있습니다.

구분	설명
whisker	상자의 좌우 또는 상하로 뻗어나간 선
박스 내부의 가로선	중앙값을 나타냅니다.
lower whisker	<ul style="list-style-type: none"> <li>최소값</li> <li>'중앙값 - 1.5 × IQR'보다 큰 데이터 중 가장 작은 값</li> </ul>
upper whisker	<ul style="list-style-type: none"> <li>최대값</li> <li>'중앙값 + 1.5 × IQR'보다 작은 데이터 중 가장 큰 값</li> </ul>
IQR	<ul style="list-style-type: none"> <li>Inter Quartile Range</li> <li>제3사분위수 - 제1사분위수</li> <li>실수 값 분포에서 1사분위수(Q1)와 3사분위수(Q3)를 뜻하고 이 3사분위수와 1사분위의 차이(Q3 - Q1)를 IQR(interquartile range)라고 합니다.</li> </ul>
점	<ul style="list-style-type: none"> <li>이상치(outlier; 아웃라이어) 또는 특이점</li> <li>lower whisker보다 작은 데이터 또는 upper whisker보다 큰 데이터가 여기에 해당됩니다.</li> </ul>

## ■ 분포 혹은 구성 비율 : boxplot 사용

- 앞의 두 그래프 모두 국가나 대륙의 구분 없이 해당 연도 모든 국가의 기대 수명을 종합한 분포를 보여준다.
- boxplot 함수를 이용하면 대륙별로 세분화된 분포 특성을 동시에 살펴볼 수 있다.



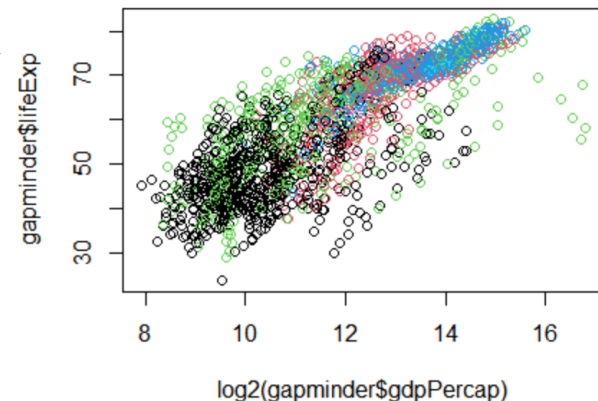
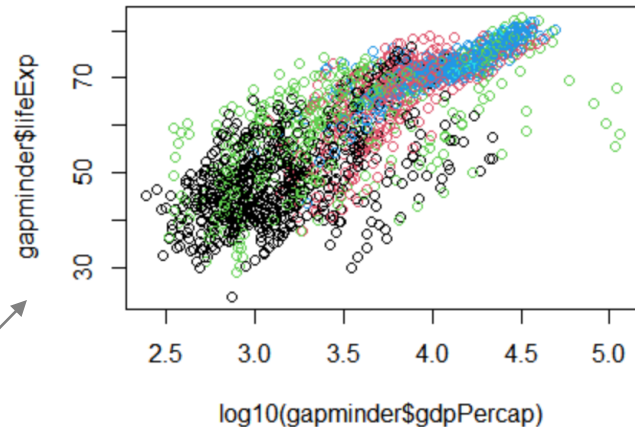
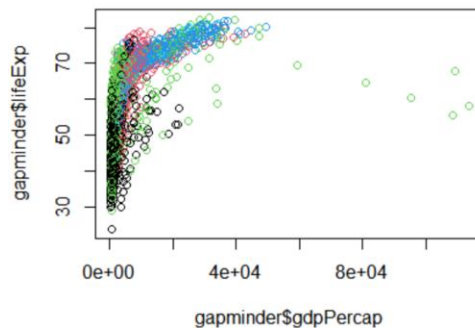
## ■ 많은 량의 데이터를 효과적으로 관찰

## ■ gapminder 데이터의 직관적 이해(3)

- ✓ gdpPercap 값의 전체 범위에 비해 낮은 범위에 샘플들이 많이 몰려 있어 관찰이 쉽지 않은 경우에는 로그 스케일(log scale)을 이용해 샘플들을 고르게 관찰 가능

```
Console C:/RSources/ ↗  
> plot(log10(gapminder$gdpPercap), gapminder$lifeExp, col =  
gapminder$continent)  
> plot(log2(gapminder$gdpPercap), gapminder$lifeExp, col =  
gapminder$continent)  
> |
```

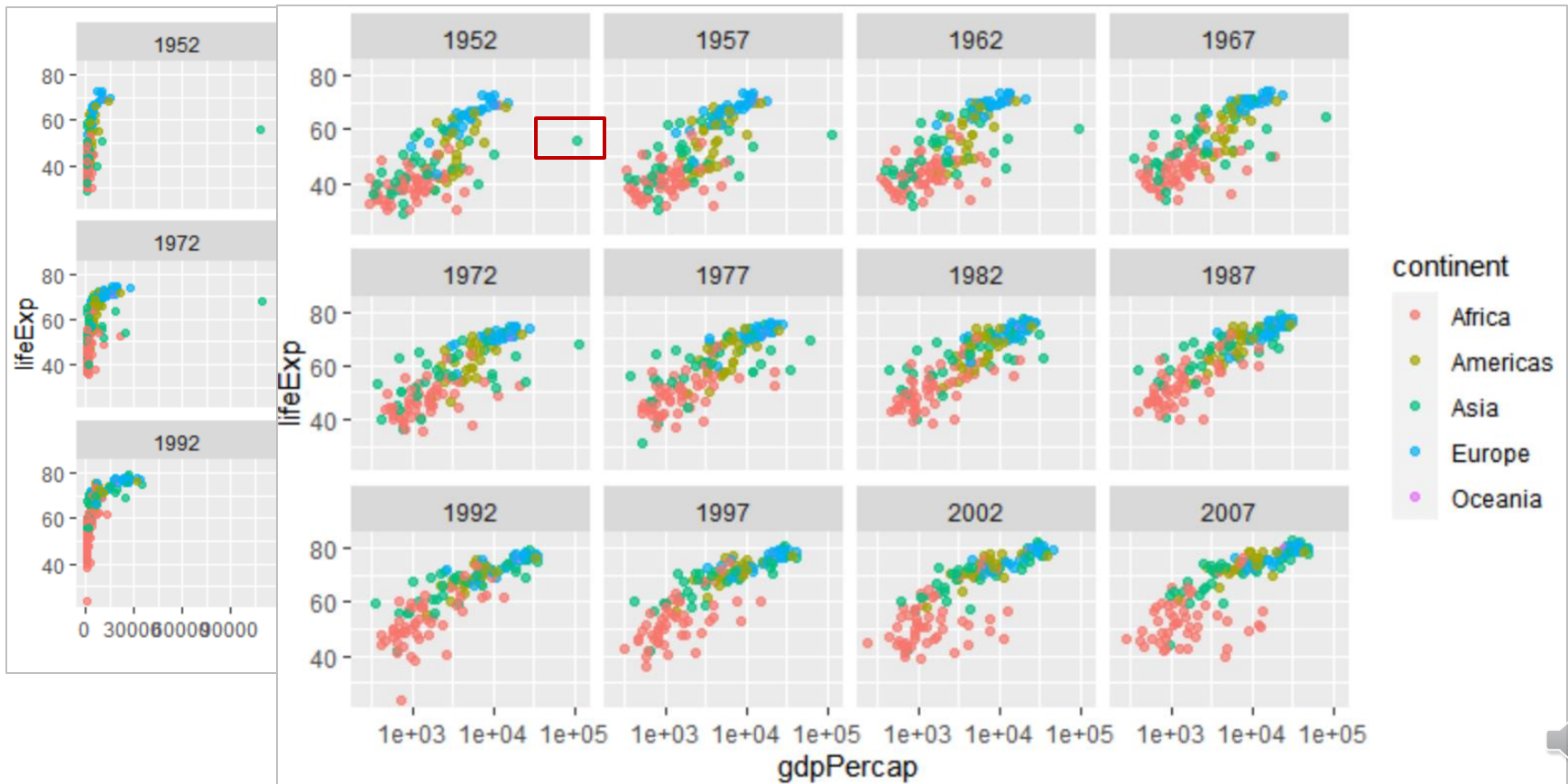
num	log2	log10
1	0.0	0.00
2	1.0	0.30
3	1.6	0.48
4	2.0	0.60
5	2.3	0.70
6	2.6	0.78
7	2.8	0.85
8	3.0	0.90
9	3.2	0.95
10	3.3	1.00
11	3.5	1.04
12	3.6	1.08
13	3.7	1.11
14	3.8	1.15
15	3.9	1.18
16	4.0	1.20
17	4.1	1.23
18	4.2	1.26
19	4.2	1.28
20	4.3	1.30





## ■ gapminder 데이터의 시각적 탐구(2)

- R을 이용해 대륙별의 경제 및 기대 수명과 이의 변화를 시각화해보기로 하자.
- `gapminder %>% ggplot(aes(gdpPercap, lifeExp, col = continent)) + geom_point(alpha = 0.7) + facet_wrap(~year) + scale_x_log10()`



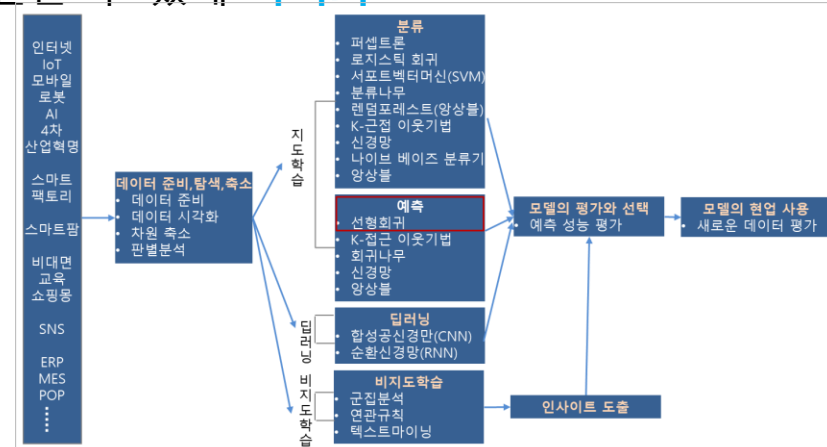
## ■ 지금까지는

- 데이터를 원하는 형태로 조작하거나 눈으로 확인할 수 있게 **가시화**
- 즉, 데이터를 다른 형태로 **'변환'**

## ■ 이제부터는

- 데이터를 가장 잘 설명하는 **모델**을 만들기
- **새로운 샘플(미래)**에 대해 **예측**
- 예) 푸드트럭 창업 (1장 4절의 예)

- 날씨, 미세먼지, 기온, 습도, 요일, 도시락 판매량을 기록한 데이터
- 모델링이란 날씨, 미세먼지, 기온, 습도, 요일(설명 변수)이 도시락 판매량(반응 변수)에 미치는 영향을 수식으로 표현하는 작업
- 모델을 사용하면 기상청의 내일 날씨 예보를 보고 도시락이 몇 개 팔릴지 예측 가능



## ■ 모델링과 예측은 데이터 과학의 핵심

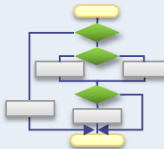

- 가장 빨리 발전하는 주제
- 경진대회는 모델링과 예측을 얼마나 정확하게 수행하는지 겨루는 장



## 7.1 모델링과 예측이란?

### ■ 모델링

- **현실 세계에서 일어나는 현상을 수학적식으로 표현**하는 행위
- 모델링을 통해 모델을 알아내고 나면, 모델을 이용하여 **새로운 사실 예측(prediction)**
- 데이터 분석 결과는 연구 대상에 대한 특징 설명 또는 어떤 값의 예측 형태로 나타남.  
이를 위해 모델링 방식을 사용함
- **모델링**이란 데이터를 발생시킨 원래 시스템을 설명하기 위해 설정한 구조
- 모델 표현 방법은 수식, 다이어그램, 알고리즘 등(데이터 사이언스 개론,p.171)

수식	다이어그램	알고리즘
가장 명쾌하나 현실적으로 많은 대상이 수식으로 설명하기 어려움	순서도 같은 형태의 처리 흐름도로 알고리즘에 비해 비교적 단순함	패턴 분석과 예측 모델 작성 등 보다 복잡한 논리적 흐름
$F = G \frac{m_1 m_2}{r^2}$ <p>중력 이론</p>		



## 7.1 모델링과 예측이란?

### ■ 간단한 예로 알아보기(영업 사원의 월급)

- 자동차 판매회사의 신입 사원인 길동은 다음과 같이 계약

조건 : 100만원 기본급에 자동차 1대 팔 때마다 90만원을 추가로 받음

- 이 조건을 기반으로 모델링
  - ✓ 판매 대수를  $x$ , 월급을  $y$ 라 하고,  $x$ 를 독립(설명) 변수  $y$ 를 종속 (반응) 변수로 간주
  - ✓ 수식으로 표현하면

모델 : 월급( $y$ ) = 1,000,000(기본급) + 900,000  $x$  (자동차 판매 대수)

- 위 수식을 **모델**이라 부름
- 급여 관련 변수를 찾고 변수 사이의 관계를 나타내는 수식을 구하는 과정을 모델링이라 부름(밤 10까지 근무, 전화 하루 100통, 고객 만남 일 10건.....)
- 모델이 있으면 **예측**이 가능
  - ✓ 다음 달에 3대를 팔면 월급이 얼마일까? → 370만원
  - ✓ 더욱 분발하여 그 다음 달에 10대를 팔면? → 1000만원



## 7.1 모델링과 예측이란?

### ■ 데이터 과학 세계의 모델링과 예측

- 데이터 사이언스 세계에서는 수집한 data로 모델링 작업
- 주어진 data를 일반적으로 훈련 데이터로부터 하나의 함수(모델)가 유추되고 나면 해당 함수에 대한 평가를 통해 파라미터를 최적화한다. 이러한 평가를 위해 교차 검증(Cross-Validation)이 이용되며 이를 위해 검증 집합을 다음의 세 가지로 나눈다.(p,343 참조, 경영을 위한 데이터 마이닝, 김종우 외 , p.88)
  1. 훈련 집합 (A Training Set) : 모델을 만드는 데 사용
  2. 검증 집합 (A Validation Set) : 처음 노형을 보정하고 일반화하는 데 사용
  3. 테스트 집합 (A Test Set) : 새로운 data 적용 시 모델 효과 추정을 위해 사용



## 7.1 모델링과 예측이란?

- 모델을 유추하기 위한 data를 **훈련 집합(training set)**라 함

- 이 데이터로부터 모델을 알아내야 함

$$X = \{x_1, x_2, x_3, \dots, x_n\}, Y = \{y_1, y_2, y_3, \dots, y_n\}$$

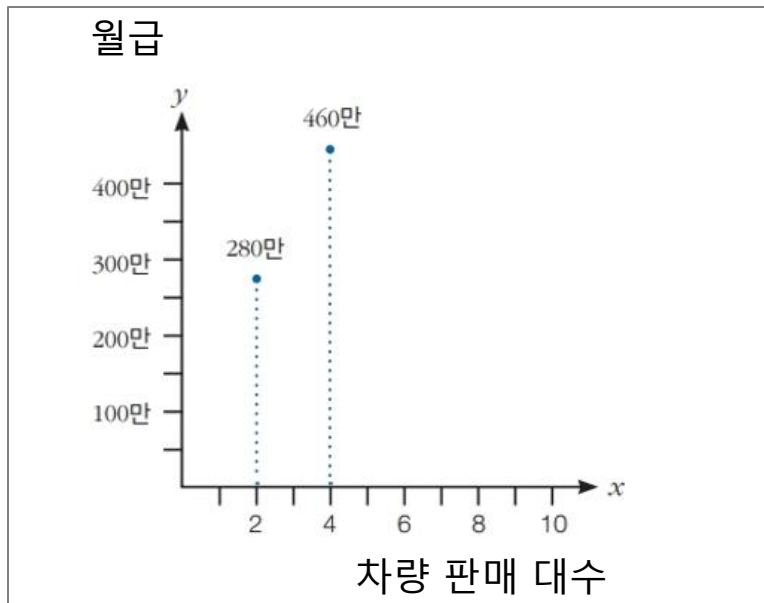
- ✓  $(x_i, y_i)$ 를  $i$ 번째 관측observation 또는  $i$ 번째 샘플sample이라 부름
- ✓ 독립 변수  $x_i$ 를 **설명 변수(explanatory variable)**, 종속 변수  $y_i$ 를 **반응 변수(response variable)**라 부름
- ✓ 또는  $x_i$ 를 특징(feature),  $y_i$ 를 레이블(label)이라 부름
- ✓ 또는  $y_i$ 를 그라운드 트루스ground truth라 부름 (GT는 정답에 해당)
- 데이터 과학에서 모델링이란 훈련 집합을 이용하여 최적의 모델을 찾는 과정



## 7.1 모델링과 예측이란?

### ■ 길동의 예에서 data로 모델 예측하기(세 번째 달 월급은?)

- 길동은 계약 내용을 제대로 모른 채 계약서에 서명
- 첫 달에 두 대를 팔아 280만원, 두 번째 달에 4대를 팔아 460만원을 받음
- 길동은 두 개의 샘플을 수집한 셈  
 $X = \{2, 4\}$ ,  $Y = \{2,800,000, 4,600,000\}$
- 훈련 집합을 그림으로 그리면



## 7.1 모델링과 예측이란?

- 길동의 예에서 data로 모델 예측하기(세 번째 달 월급은?)

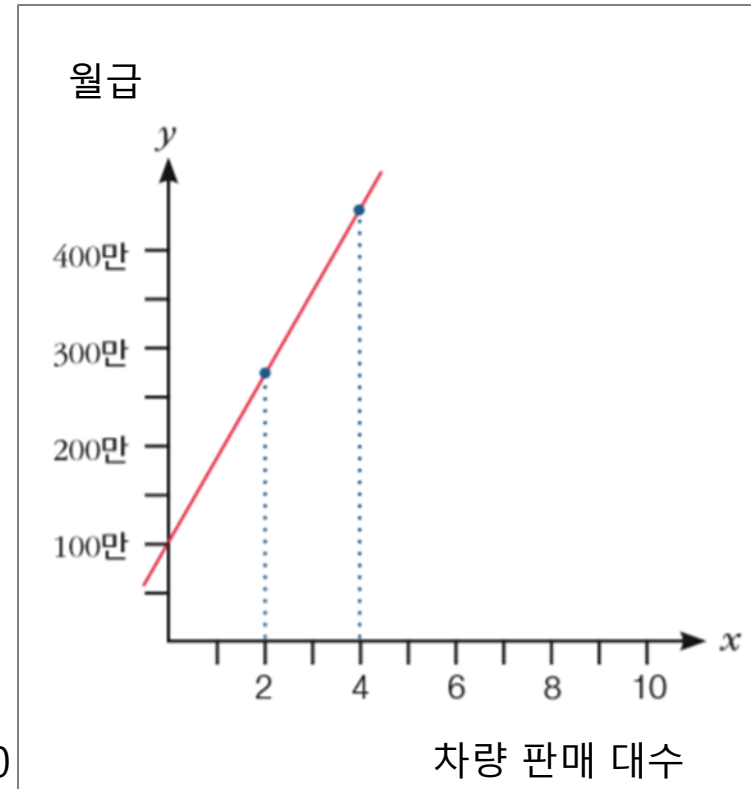
선배로부터 기본급에 판매 대수에 따라 월급이 지급된다는 사실의 정보 획득

아하 그렇다면 월급 = 기본급 + 판매 수당 x 차량 판매 대수

첫 번째 달  $2,800,000 = \text{기본급} + \text{판매 수당} \times 2$

두 번째 달  $4,600,000 = \text{기본급} + \text{판매 수당} \times 4$

두식을 풀면 : 기본급 = 1,000,000, 판매 수당 = 900,000  
따라서 모델은 :  $y = 1,000,000 + 900,000x$ (판매 대수)





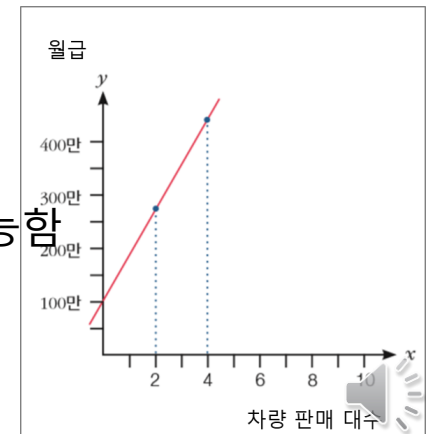
## 7.1 모델링과 예측이란?

### ■ 길동이 한 일을 정리하면,

- 정보(data)를 바탕으로 식을 수립하는 일을 **모델 선택(model selection)**이라 부름
  - ✓ 만일 1대당 90만원의 고정 인센티브제가 아니라면 다른 방정식을 선택해야 함
  - ✓ 3대까지는 수당이 없고(기본급으로 대체 4대부터 판매 대수에 따라 누진 적용)
- 수학에서는  $y = \alpha_0 + \alpha_1 x$  에서  $\alpha_0, \alpha_1$  를 계수, 데이터 과학은 매개변수(parameter)라 부름
- 훈련 집합을 가장 잘 설명하는 최적의 매개변수 값을 알아내는 과정을 모델 적합(model fitting)이라 부름. 학습(learning) 또는 훈련(training)이라고도 부름
- 모델 적합으로 모델을 알아내면 예측이 가능

### ■ 예측

- 모델  $y = 900000x + 1000000$  을 가지고 새로운 샘플에 대해서 예측이 가능함
- 예)  $x=5$  (즉 5대를 팔면),  $y = 900000 * 5 + 1000000 = 550$ 만 원

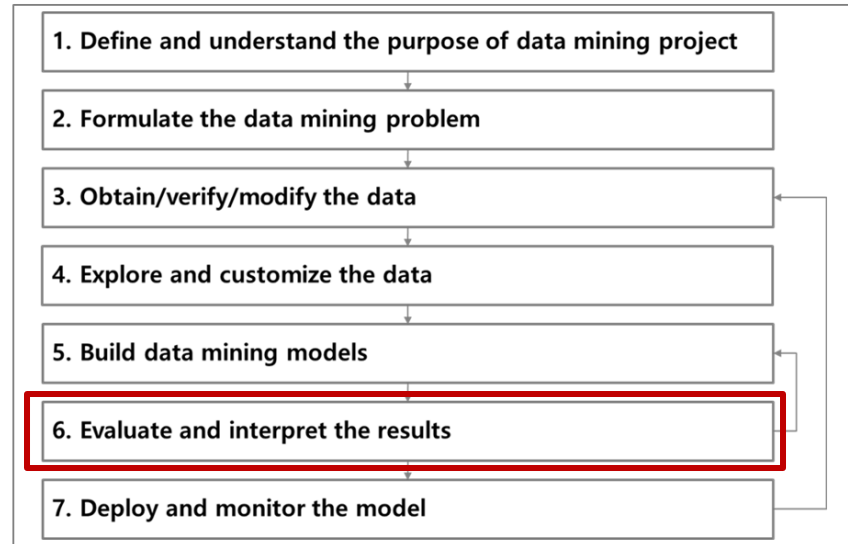


## 7.1 모델링과 예측이란?

### ■ 모델의 품질 평가

- 모델이 범하는 오류로 평가함
- 이 예는 불확실성이 없는 월급의 예이므로 오류가 0
- 주어진 data를 일반적으로 훈련 데이터로부터 하나의 함수(모델)가 유추되고 나면 해당 함수에 대한 평가를 통해 매개변수 (parameter)를 최적화한다. 이러한 평가를 위해 교차 검증(Cross-Validation)이 이용되며 이를 위해 검증 단계를 거침

### 데이터 마이닝 Process



## 7.2 현실 세계의 모델링

### ■ 데이터 과학이 다루는 문제는 불확실성이 개입

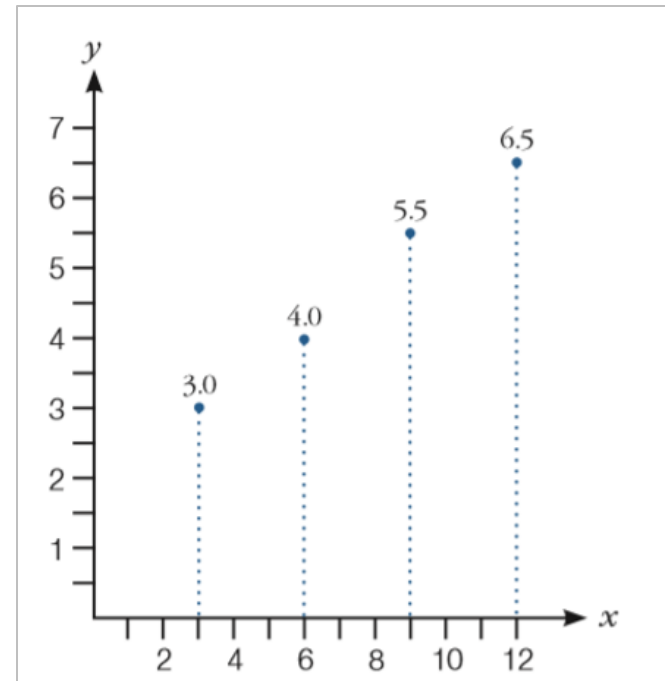
- 같은 시간에 같은 사람의 체온을 측정하면 36.0, 36.3, 36.7도와 같이 다르게 나타남
- 또한 혈압도 110, 115 심지어 계단을 오른 후는 120이 넘기도 한다.
- 현실 세계에서 발생하는 data는 환경과 측정 기계 따라 상당한 차이를 나타냄

### ■ 자극과 반응 물리 실험 예

- 전기 자극에 따른 물체의 이동 거리 실험
- 전기량  $x$ , 물체의 이동 거리  $y$ 로 하고 네 번의 실험을 통해 데이터 수집

$$X = \{3.0, 6.0, 9.0, 12.0\},$$

$$Y = \{3.0, 4.0, 5.5, 6.5\}$$

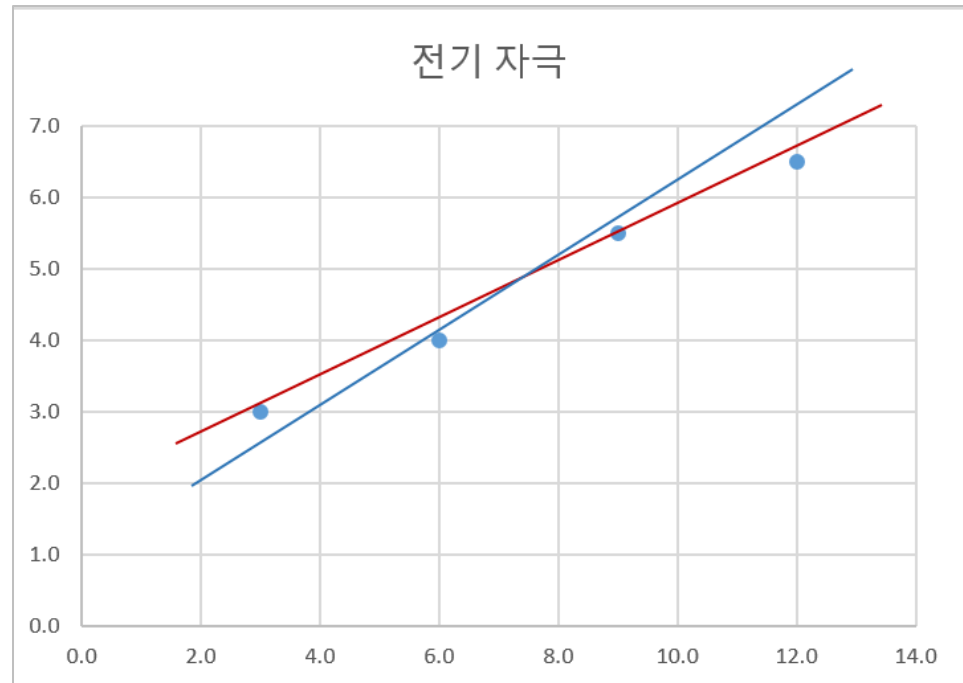


## 7.2 현실 세계의 모델링

### ■ 자극과 반응 물리 실험 예

- 대략 선형을 이루는 것을 확인하고, 선형 방정식을 사용하기로 결정 (모델 선택)
- 샘플 4개를 각 방정식에 대입하면, 훈련 집합이 선형이 아님

	1차	2차	3차	4차
X	3.0	6.0	9.0	12.0
Y	3.0	4.0	5.5	6.5



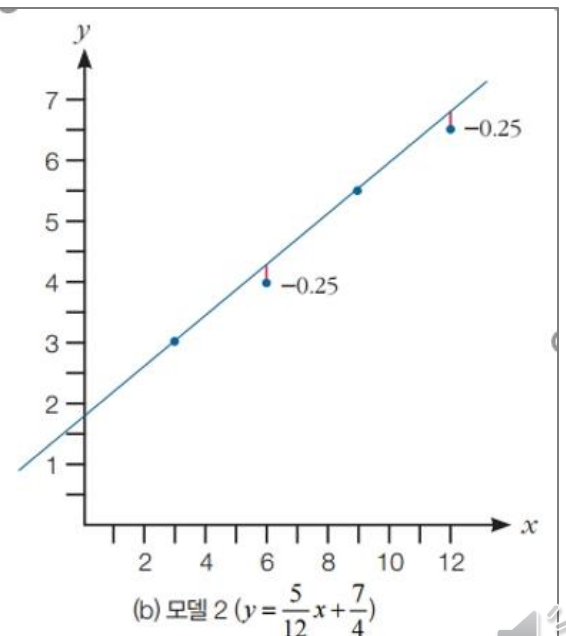
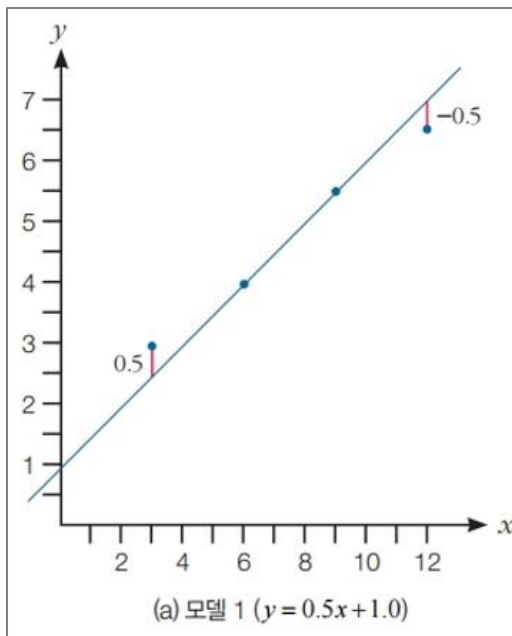
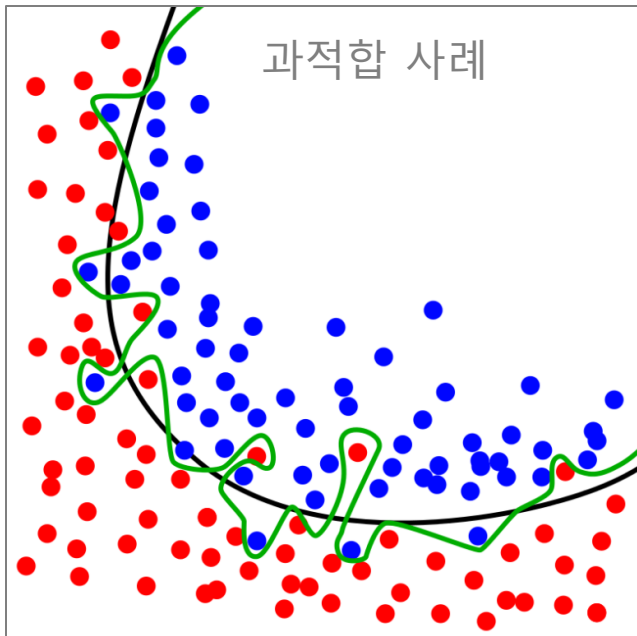
- 모델 선택을 실패한 셈. 어떻게 해야 할까?
  - 모델로 선형을 버리고 2차 또는 3차 같은 고차 방정식을 사용?
  - 선형 모델을 사용하되 오차를 허용?



## 7.2 현실 세계의 모델링

### ■ 자극과 반응 물리 실험 예

- 복잡한 모델은 과잉적합(overfitting)의 위험 (P,291)
- 과잉 적합 : 일반화 능력을 상실하는 현상
- 현실 세계의 데이터에서 오차 0은 불가능 → 오차를 허용하고 선형 방정식을 사용



## 7.2 현실 세계의 모델링

## ■ 성능 분석

- [그림(a)]의 모델  $y=0.5x+1.0$ 의 오차 분석

x1	3.0	6.0	9.0	12.0
예측값 $f(x_i)$	2.5	4.0	5.5	7.0
그라운드 투루스 $y_i$	3.0	4.0	5.5	6.5
오차	0.5	0.0	0.0	-0.5

- 평균 제곱 오차(MSE, Mean squared error)

$$E = \frac{1}{4}((0.5)^2 + (0.0)^2 + (0.0)^2 + (-0.5)^2) = 0.125$$

- MSE를 일반화하면 :  $E = \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$  식 (3)



## 7.2 현실 세계의 모델링

## ■ 성능 분석

- [그림(b)]의 모델  $y=5/12x+7/4$ 의 오차 분석

x1	3.0	6.0	9.0	12.0
예측값 $f(x_i)$	3.0	4.25	5.5	6.75
그라운드 투루스 $y_i$	3.0	4.0	5.5	6.5
오차	0.0	-0.25	0.0	-0.25

$$E = \frac{1}{4}((0.0)^2 + (-0.25)^2 + (0.0)^2 + (-0.25)^2) = 0.03125$$

- 두 모델 중에 어느 것이 좋은가?  
→ MSE가 낮은 두 번째 모델이 우월함
- 이 훈련 집합을 보다 잘 설명하는 (즉 MSE가 더 낮은) 더 좋은 모델이 있는가?  
→ 모델링은 최적화 문제(optimization problem)



# Thank you

