

머신러닝과 빅데이터분석(R)

12주차 선형회귀



박길식 교수



고려사이버대학교
THE CYBER UNIVERSITY OF KOREA



학습 목표

-  선형 회귀 분석의 개념을 설명할 수 있다.
-  선형 회귀 예측 모델을 구현할 수 있다.



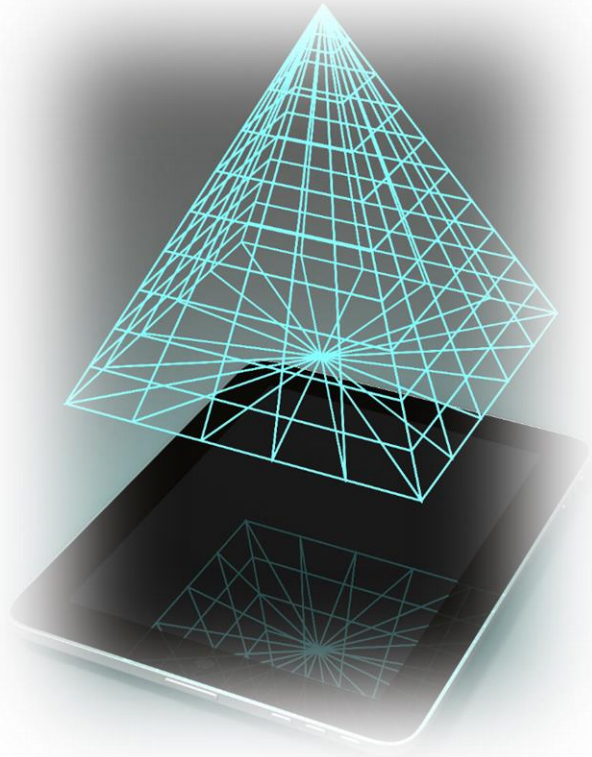
학습 목차

- 1 단순 선형 회귀와 다중 선형 회귀의 분석
- 2 선형 회귀 실습

CHAPTER

단순 선형 회귀와 다중 선형 회귀의 분석

모델링과 예측은 데이터 과학의 핵심



모델링(Modeling)

현실 세계에서 일어나는 현상을
수학식으로 표현하는 행위

- » 모델을 통해 모델을 구성하고,
모델을 이용하여 새로운 사실을
예측(Prediction)할 수 있음

[01] 단순 선형 회귀 모델



“영업 사원의 월급”

㉞ 자동차 판매회사의 신입 사원인 홍길동은 다음과 같이 계약

- 조건 : 100만 원 기본급에 자동차 1대 팔 때마다 90만 원을 추가로 받음

㉞ 이 조건을 기반으로 모델링

- 판매 대수를 x , 월급을 y 라 하고 x 를 독립변수, y 를 종속변수로 간주
- 수식으로 표현하면

- 모델 : $y = 900,000x + 1,000,000$

[01] 단순 선형 회귀 모델



“영업 사원의 월급”

② 모델링

- 영업사원 홍길동이 변수를 뽑고 변수 사이의 관계를 나타내는 수식을 구하는 과정

③ 모델이 있으면 예측이 가능

- 다음 달에 3대를 팔면 월급이 얼마일까?
➡ 370만 원
- 더욱 분발하여 다다음 달에 20대를 팔면?
➡ 1,900만 원

[01] 단순 선형 회귀 모델



“영업 사원의 월급”

- ⌘ 주어진 데이터로부터 모델을 알아내야 함
- ⌘ **훈련 집합(Training set)** : 주어진 데이터(X 와 Y)
 - (x_i, y_i) 를 i 번째 관측(Observation) 또는 i 번째 샘플(Sample)이라 부름
 - 독립변수 x_i : 설명 변수 또는 특징 벡터
 - 종속변수 y_i : 반응 변수, 레이블(Label) 또는 목표값(Ground Truth)

데이터 과학에서 모델링

- 훈련 집합을 이용하여 최적의 모델을 구성하는 과정
 - ➔ 훈련 집합을 가장 잘 설명하는 모델을 구성하는 과정

[01] 단순 선형 회귀 모델



모델링의 예

- ⊕ 홍길동은 계약 내용을 제대로 모른 채 계약서에 서명하는 실수를 범함
- ⊕ 첫 달에 2대를 팔아 280만 원,
두 번째 달에 4대를 팔아 460만 원을 받음
- ⊕ 홍길동은 두 개의 샘플을 수집한다고 볼 수 있음

$$X = \{2, 4\}, \quad Y = \{2,800,000, 4,600,000\}$$

[01] 단순 선형 회귀 모델



모델링의 예

- ㉔ 홍길동은 선배로부터 기본급과 판매 대수에 비례한 인센티브를 더해 월급을 받는다고 사실을 알게 되어 선형 방정식을 세움

$$Y = \alpha X + \beta$$

- ㉔ β 는 기본급, α 는 1대 팔 때마다 받는 인센티브에 해당

$$2,800,000 = 2\alpha + \beta$$

$$4,600,000 = 4\alpha + \beta$$

- ㉔ 두 식을 풀면, $\beta=900,000$, $\alpha=1,000,000$

- ㉔ 즉, $y = 900,000x + 1,000,000 \leftarrow$ 모델링

[01] 단순 선형 회귀 모델



모델링의 예

- ② 모델링(Modeling) 또는 모델 구성(Model Construction)
: 선배로부터 받은 정보를 바탕으로 식을 수립하는 일
 - 만일 1대당 90만 원의 고정 인센티브제가 아니라면
다른 방정식을 수립해야 함
- ② 수학적 구조에서는 α 과 β 를 계수, 데이터 과학에서는 매개변수(Parameter)
 - 대부분의 머신러닝은 모델을 구성하였을 때
매개변수의 값을 파악하기가 어려움(Black-Box)

[01] 단순 선형 회귀 모델



모델링의 예

- ② 모델 적합(Model Fitting), 학습(Learning), 또는 훈련(Training)
 - : 훈련 집합을 가장 잘 설명하는 최적의 매개변수 값을 알아내는 과정
 - 복잡한 모델은 과적합(Overfitting)의 위험을 가지고 있음
 - 현실 세계의 데이터에서 오차 0은 불가능
 - ➡ 오차를 어느 정도 허용하고 모델 구성

[01] 단순 선형 회귀 모델



모델링의 예 : 홍길동의 일 정리

» 예측

- 모델 적합 후 새로운 데이터(Unseen Data) 또는 샘플이 주어지면 예측이 가능
- 모델($y = 900,000x + 1,000,000$)을 이용하여 훈련 집합에 없는 새로운 샘플에 관해 예측이 가능함

이
예

$x = 5$ (즉, 5대를 팔면),

$y = 900,000 * 5 + 1,000,000 = 550$ 만 원

— [01] 단순 선형 회귀 모델



모델링의 예 : 홍길동의 일 정리

⌘ 모델의 성능 평가

- ─ 모델이 도출하는 예측값과 관측값을 비교하여 오류를 평가함
- ─ 이 예는 불확실성이 없는 월급의 예이므로 오류가 0

[01] 단순 선형 회귀 모델



회귀(Regression)

- ② 통계학에서 유래된 용어
- ② 독립변수가 변할 때, 종속변수가 어떻게 변하는지를 수식으로 표현하는 과정

단순 회귀
(Simple Regression)

독립변수의 수가
하나인 경우

다중 회귀
(Multiple Regression)

독립변수의 수가
2개 이상인 경우

[01] 단순 선형 회귀 모델



회귀(Regression)

- ㉔ 앞서 홍길동의 월급 계산처럼 훈련 집합을 이용하여 모델을 구성하는 과정
 - ➡ “회귀 문제를 푼다” 또는 “회귀 분석을 한다”라고 표현
- ㉔ 앞서 홍길동의 월급 계산은 단순선형 회귀모델을 구성한다고 볼 수 있음

[02] 다중 선형 회귀 모델

 현실 세계의 데이터는 설명(독립) 변수가 여러 개

② 월급에 영향을 미치는 변수

➡ 판매 대수뿐만 아니라 근무 연수, 직급 등

② 제동 거리에 영향을 미치는 변수

➡ 속도뿐만 아니라 날씨나 브레이크의 종류 등

② 일반적으로 표시하면

독립변수

x, u, v, w, z, \dots

종속변수

y

[02] 다중 선형 회귀 모델



다중 선형 회귀(Multiple Linear Regression)

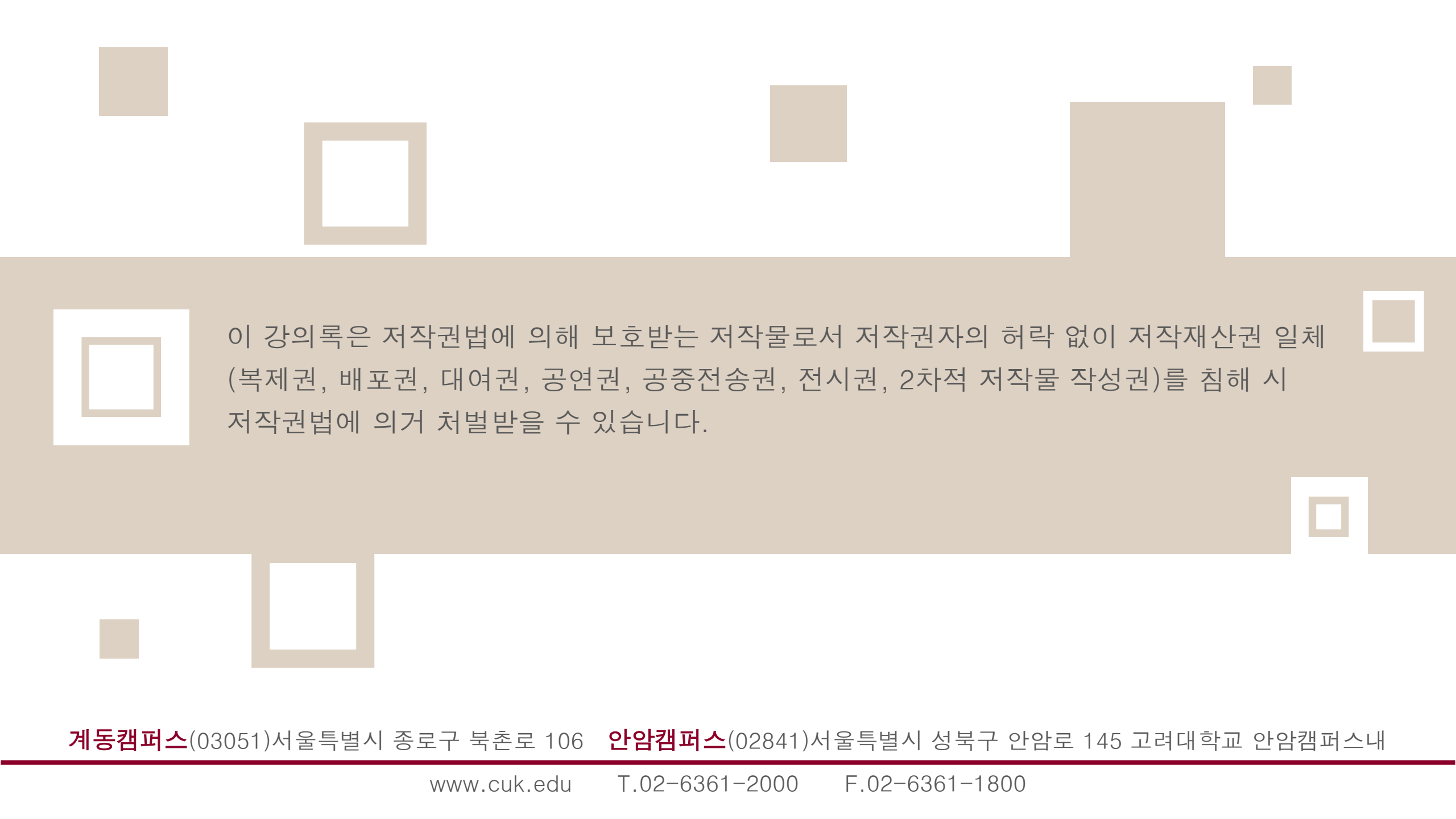
⊕ 설명(독립) 변수가 2개 이상인 선형 회귀

$$Y = \alpha_1 x + \alpha_2 u + \alpha_3 v + \cdots + \beta$$

⊕ 설명(독립) 변수가 2개인 경우에는 매개변수가 3개

⊕ 일반적으로 설명(독립) 변수가 k개이면 매개변수는 k+1개

⊕ 다중선형 회귀분석도 단순선형 회귀분석과 마찬가지로 $\ln()$ 함수로 구성할 수 있음



이 강의록은 저작권법에 의해 보호받는 저작물로서 저작권자의 허락 없이 저작권재산권 일체 (복제권, 배포권, 대여권, 공연권, 공중전송권, 전시권, 2차적 저작물 작성권)를 침해 시 저작권법에 의거 처벌받을 수 있습니다.