Python을 활용한 데이터 분석 강의

기초통계방법론

HYUNDAI MOBIS

통계의 기초

- 모집단(population) : 연구 대상 전부
- 표본(sample) : 자료에 담겨 있는 모집단의 일부분

ex) 2008년 한국종합사회조사(Korean General Social Survey: KGSS)

모집단 : 한국성인남녀 모두

표본 : 1,508명의 한국성인남녀

기술통계와 추리통계

- 기술통계(descriptive statictics)
  - 자료(표본)의 정보를 요약
  - 통계량(statistic) : 숫자로 표현한 자료(표본)의 요약

TABLE 2.1 Descriptive Statistics for Donut and Weight Data

Variable	Observations (*N*)	Mean	Standard deviation	Minimum	Maximum
Weight	13	171.85	76.16	70	310
Donuts	13	5.41	6.85	0	20.5

기술통계와 추리통계

- 추리통계(inferential statictics)
  - 자료(표본)의 정보에 기반하여 모집단의 속성을 예측(추론)함
  - 모수(parameter) : 숫자로 표현한 모집단의 요약
  - 모수 = 통계량 + 불확실성

HYUNDAI MOBIS

변수와 변수의 측정

- 변수(variable)
  - 표본 혹은 모집단 안의 개체 또는 관찰값들이 갖는(서로 다른 값을 취하는) 속성

- 양적 변수와 질적 변수
  - 양적 변수(quantitative variable) : 변수의 값이 숫자로 표현됨
    - ex. 나이, 연봉
  - 질적 변수(qualitative variable) : 변수의 값이 범주로 표현됨
    - ex. 종교, 학점(A,B,C..)

보통 자료에서 질적 변수의 각 범주는 숫자로 표현

(ex. 예:1, 아니오:0)

변수와 변수의 측정

- 이산형 변수와 연속형 변수
  - 이산형 변수(discrete variable) : 변수의 값이 정수로 표현됨
    - ex. 형제 자매의 수(0,1,2,3...)
  - 연속형 변수(continuous variable) : 변수가 가질 수 있는 값이 무한함
    - ex. 연봉

측정 척도에 따른 변수의 구분

- 명목 척도 변수(nominal scale variable)
  - 변수값이 뚜렷한 순서가 없는 범주의 나열
    - ex. 종교

- 순서 척도 변수(ordinal scale variable)
  - 변수값들 간에 쉽게 이해 가능한 순서가 존재
    - ex. 사회적 지위(상류층, 중산층 하류층)

- 등간 척도 변수(interval scale variable)
  - 변수값들 간에 순서가 있고 각 값들 간의 간격이 일정
    - ex. 나이

집중경향치

[1] 평균 (M, $\overline{X}$)

- 분포에 있어 평균치는 점수의 합을 점수의 개수로 나눈 것 $\mu = \frac{\sum X}{N}$

[2] 최빈값 (Mo)

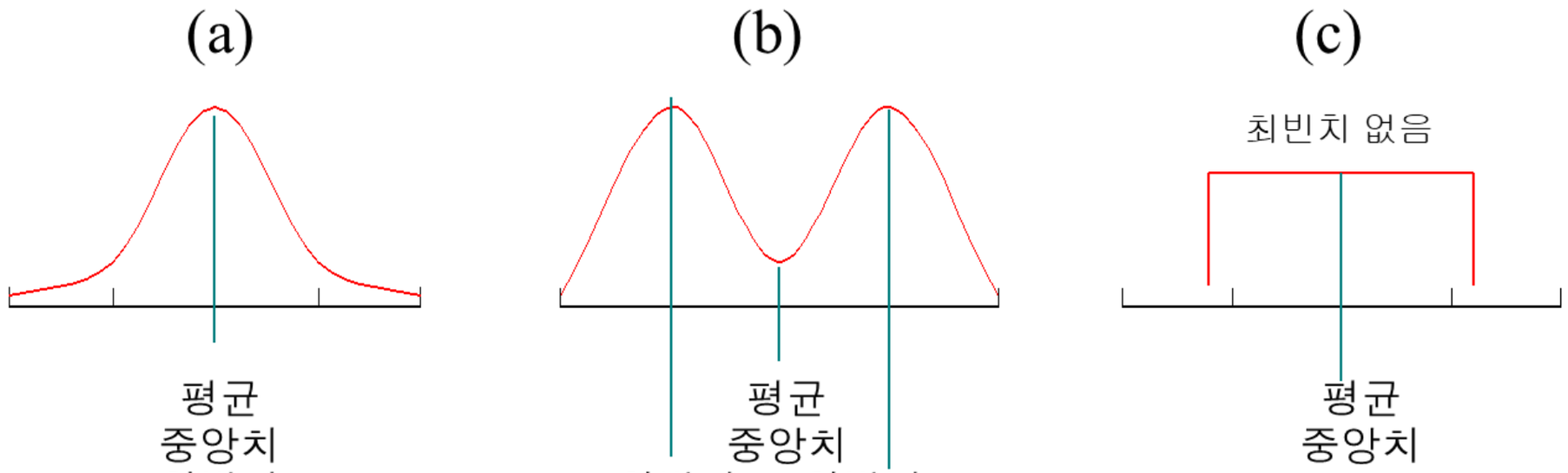
- 가장 최대의 빈도를 갖는 점수나 유목
- 어떤 빈도가 아닌 점수나 범주

[3] 중앙값 (Mdn)

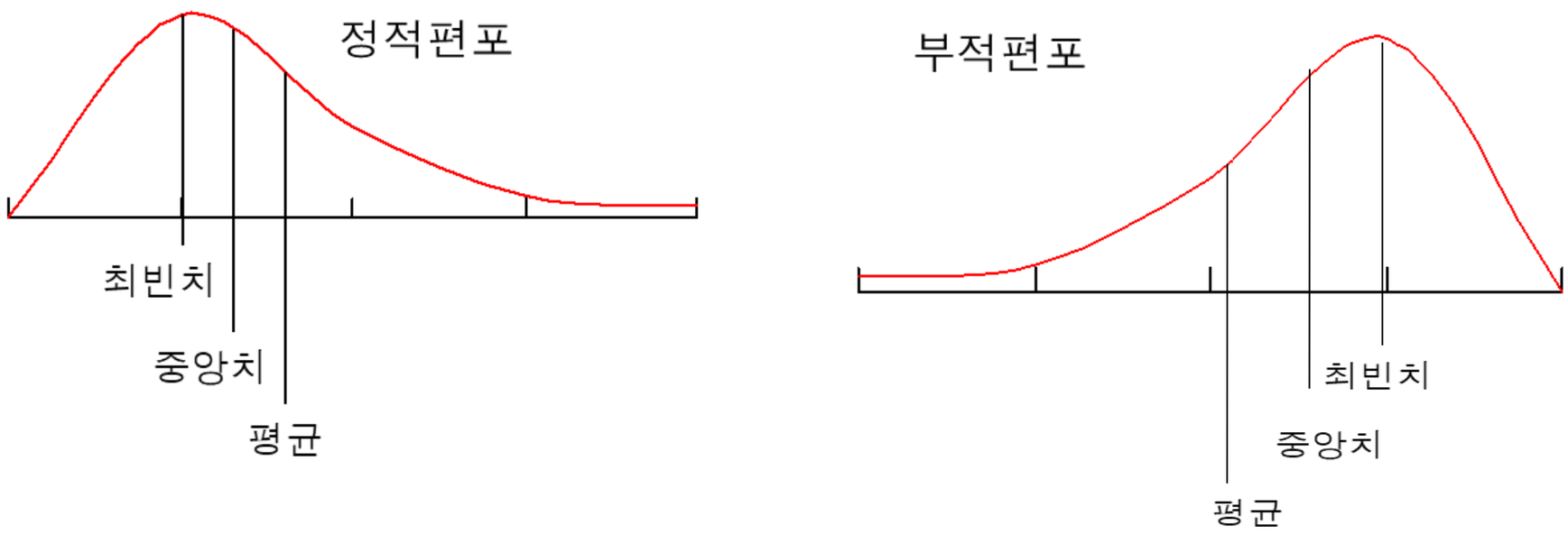
- 측정치를 크기의 순서로 배열해 놓았을 때 정확히 절반으로 나누는 값
- 50번째 백분위 점수

집중경향치 분포

- 대칭분포



- 비대칭 분포



표준편차

- 변산도(variability)

  - 분포에 있는 점수들이 흩어져 있는지, 아니면 함께 몰려 있는지의 정도를 양적으로 나타내는 것

- 표준편차 (standard deviation)

  - 자료가 평균을 중심으로 얼마나 퍼져 있는지를 나타내는 대표적인 수치

$$\sigma_X = \sqrt{\mathrm{V}(X)}$$

$$\mathrm{V}(X) = \mathrm{E}((X - \mu)^2)$$

공분산

- 2개의 확률변수의 상관정도를 나타내는 값
- $$\operatorname{Cov}(X,Y) \equiv \operatorname{E}[(X - \mu_X)(Y - \mu_Y)],$$

$$\operatorname{Cov}(X, X) = \operatorname{Var}(X)$$

$$\operatorname{Cov}(X, Y) = \operatorname{Cov}(Y, X)$$

- 공분산은 X와 Y의 단위에 영향을 받음
  단위에 영향을 받지 않기 위해서 "상관계수"사용

상관계수와 결정계수

- **상관계수 r**

  - 독립변수와 종속변수간의 선형적인 관계를 나타내는 척도

  - [-1 , 1]

  - 단순 관련성이 아닌 선형적인 관계

$$R = \frac{\mathrm{cov}(x, y)}{\sqrt{\mathrm{var}(x)\mathrm{var}(y)}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2(y_i - \bar{y})^2}}$$

상관계수와 결정계수

- **결정계수 R^2**

  - 선형회귀분석에서 회귀직선의 적합도를 평가하거나 종속변수에 대한 설명변수들의 설명력을 알고자 할 때

  - **R^2** = 0.45 일때, y의 변동은 x의 변동에 의해 45% 정도 설명된다.

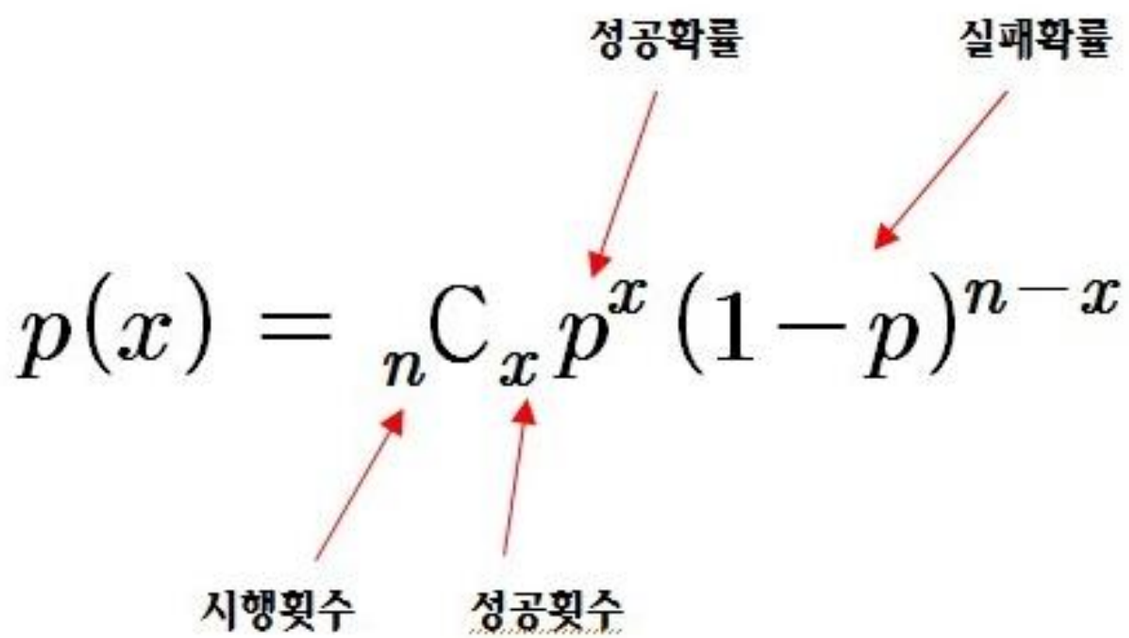
  - [0, 1]

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})}{\sum(y_i - \bar{y})^2} = \frac{\textit{회귀선에 의해 설명되는 변동}}{\textit{전체 변동}}$$

이산확률분포 – 이항분포

- **이항분포**

  - 어떤 시행에서 사건 A가 일어날 확률 = p
  - 이 시행을 독립적으로 n회 반복
  - 그 중에서 x회만 A가 일어날 확률



이산확률분포 – 포아송분포

- **포아송분포**

  - 단위 시간 안에 어떤 사건이 몇 번 발생할 것인지를 표현하는 이산 확률 분포

  - Ex> - 어느 주말 일요일 서울에서 발생한 교통사고 사망자의 수

    - 어느 보험 회사의 주말 동안의 보험 클레임 수

    - 어느 하루 동안 지정된 생산라인에서 발생한 불량품의 개수

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \qquad x = 0,1,\ldots, \ \lambda > 0$$

λ : 단위 시간 또는 단위 공간 내의 발생횟수의 평균

HYUNDAI MOBIS

회귀분석

HYUNDAI MOBIS

회귀분석

- 두 (또는 그 이 상) 양적 변수들의 관계를 기술(description)하고 추론(inference)하는 통계 방법
  - 단순회귀분석 : 하나의 종속변수와 하나의 독립변수
  - 다중회귀분석 : 하나의 종속변수와 여러 개의 독립변수

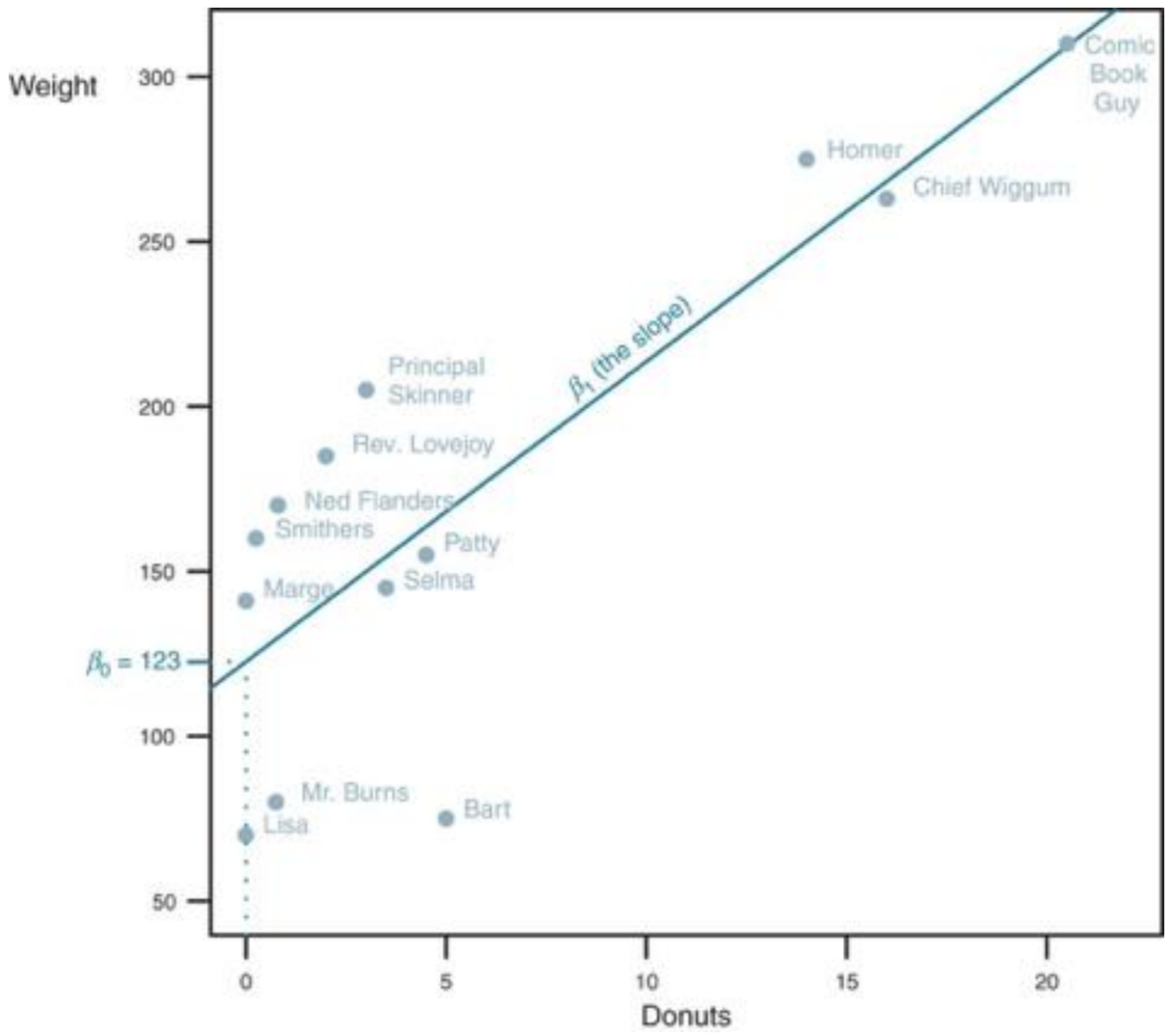
HYUNDAI MOBIS

단순회귀분석

$$Y = b_0 + b_1 * X + e$$

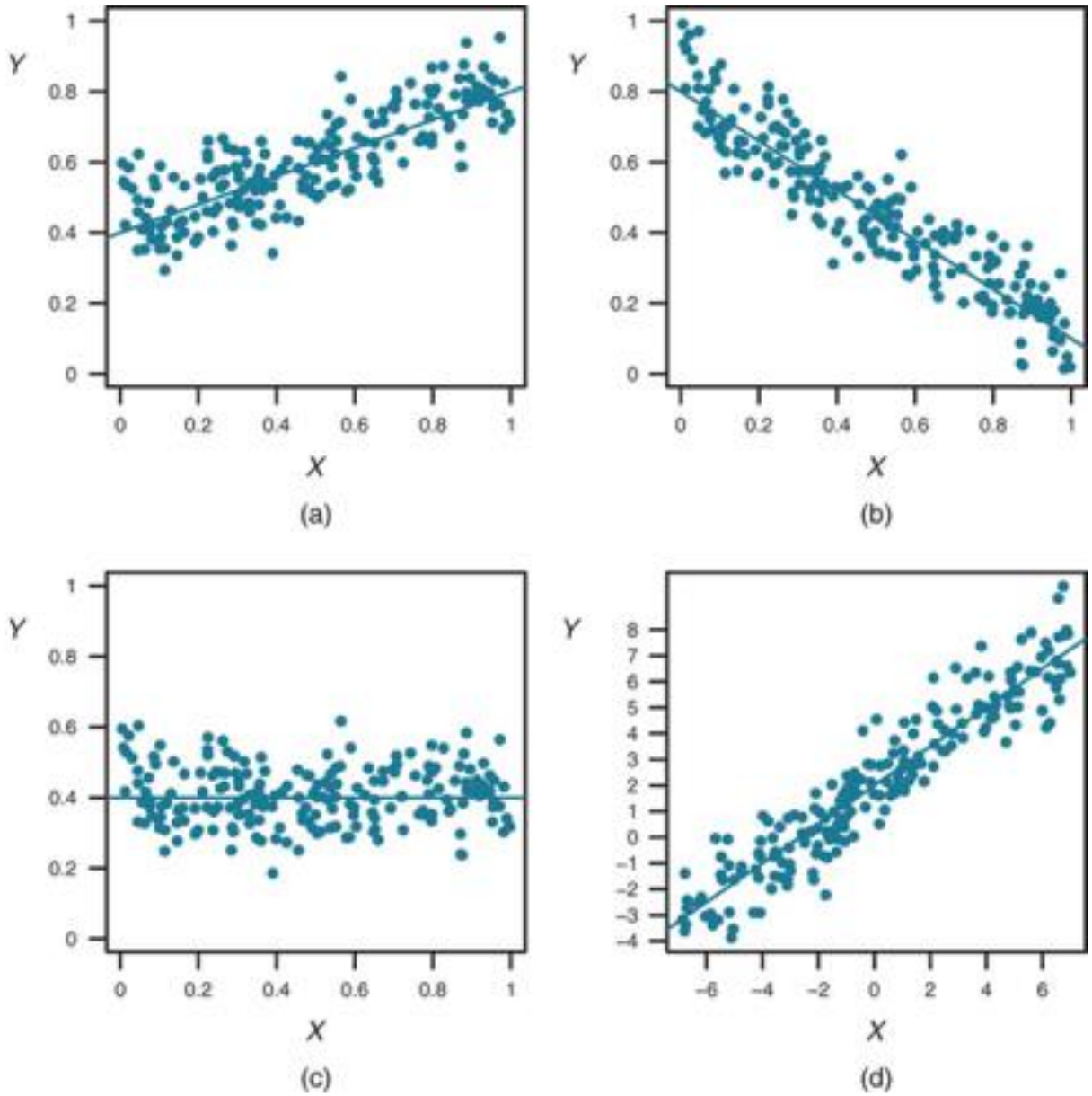
- Y(종속변수)는 X(독립변수)의 선형 함수(linear function)
- b_1 : 기울기(slope; X가 한 단위 증가할 때 생기는 Y의 변화), 회귀계수(regression coefficient)라고 부른다
- b_0 : 절편(intercept or constant; X=0일 때 Y의 값)
- e : X를 제외하고 Y에 영향을 주는 요인들

단순회귀분석

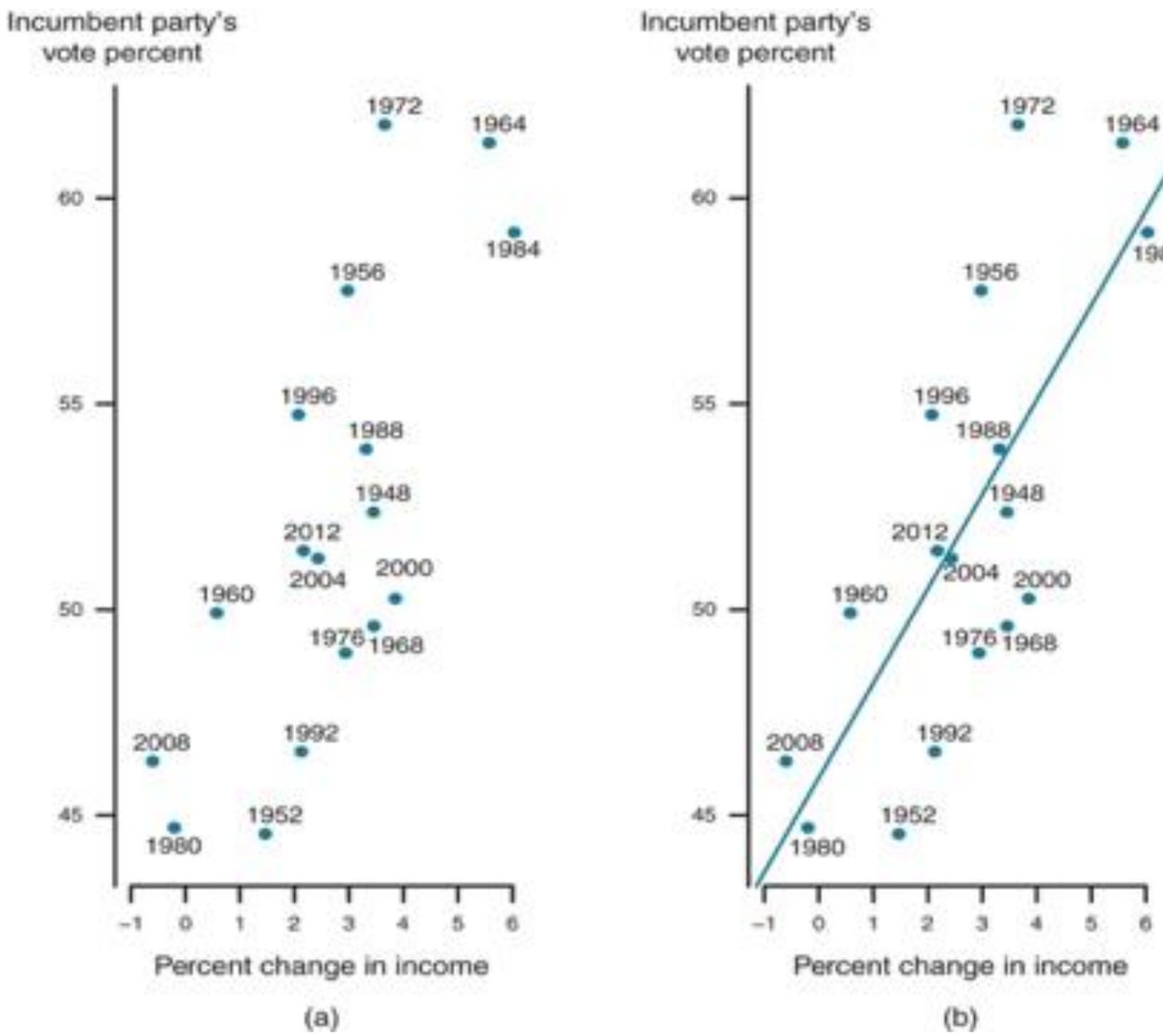


HYUNDAI MOBIS

단순회귀분석



단순회귀분석: 회귀선 찾기



최소제곱법(Least Square)

$$Y = \beta_0 + \beta_1 * X + e$$

(모집단 수준의 회귀선)

- 어떻게 자료에서 β_0와 β_1를 추정할 수 있을까?
- b_0와 b_1을 β_0와 β_1의 추정값이라고 본다

$$\hat{Y}_i = \widehat{\beta_0} + \widehat{\beta_1} X_i$$

$$\hat{Y}_i = b_0 + b_1 X_i$$

$$\min_{b_0, b_1} \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2$$

HYUNDAI MOBIS

회귀모형의 적합도

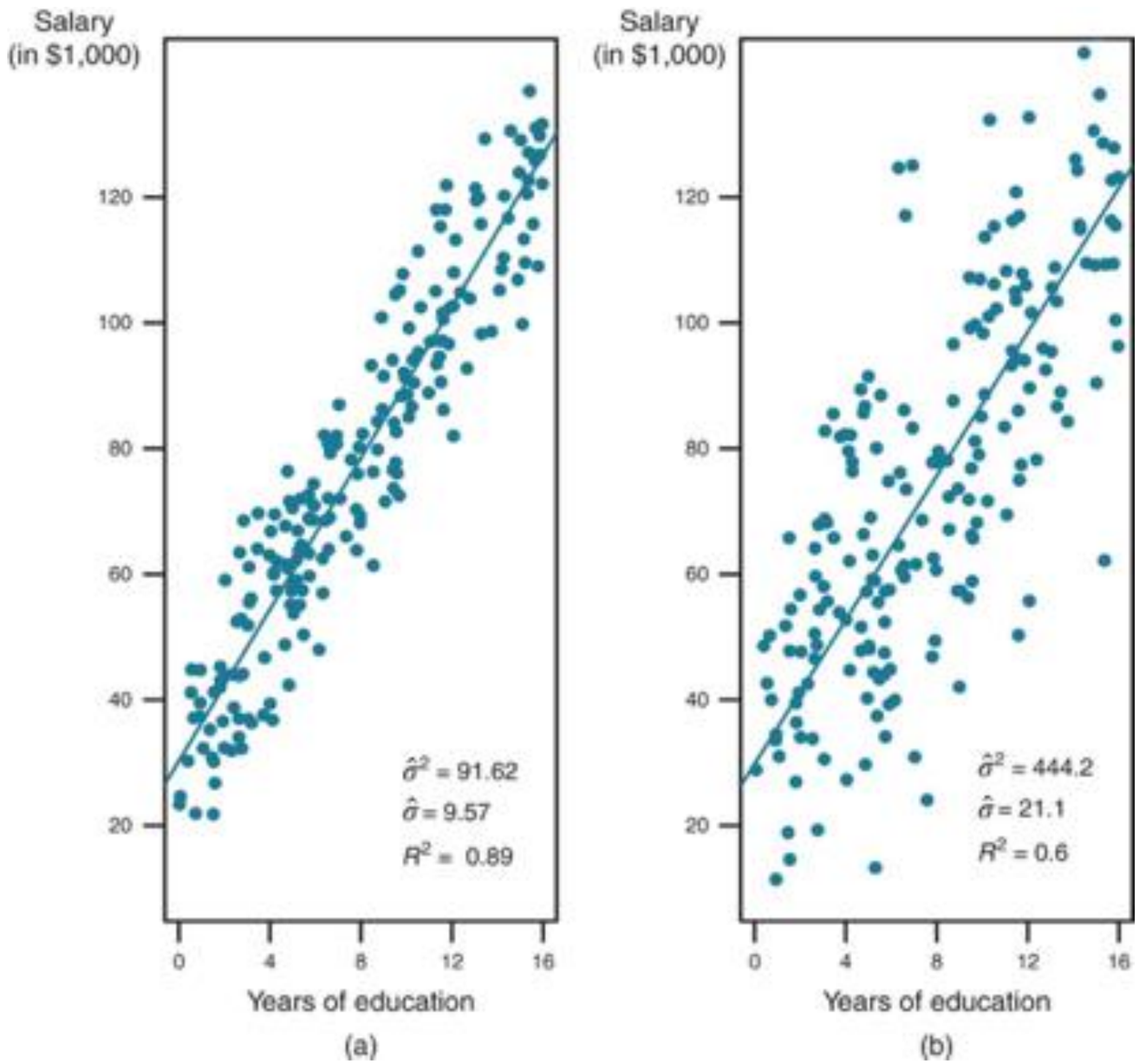
- 찾아낸 회귀선이 얼마나 자료를 잘 설명하는가(즉 회귀선이 얼마나 자료에 적합한가)를 보는 것

- R^2 : 독립변수에 의해 설명되어지는 종속변수의 분산의 비율(단순회귀분석에서는 X와 Y간의 상관계수의 제곱)

- 회귀의 표준 오차(SER: standard error of the regression) : 종속변수에 존재하는 회귀 잔차의 크기

HYUNDAI MOBIS

회귀모형의 적합도


Salary
(in $1,000)
120
100
80
60
40
20
0
4
8
12
16
Years of education
$\hat{\sigma}^2 = 91.62$
$\hat{\sigma} = 9.57$
$R^2 = 0.89$
(a)
Salary
(in $1,000)
120
100
80
60
40
20
0
4
8
12
16
Years of education
$\hat{\sigma}^2 = 444.2$
$\hat{\sigma} = 21.1$
$R^2 = 0.6$
(b)

회귀분석의 절차

- 종속변수와 독립변수를 설정
- 회귀분석을 진행
- 회귀계수와 그에 해당하는 표준 오차를 찾아 독립변수가 통계적으로 의미 있는 요인인지를 해석함(t-값 혹은 p-값을 본다)
- R^2를 보고 모델의 적합도를 해석

HYUNDAI MOBIS

다중회귀분석

- 하나의 종속변수(Y)를 설명할 수 있는 요인들을 많음

- 하나의 독립변수(X_1)가 종속변수(Y)에 주는 효과를 다른 독립변수들(X_2, X_3, X_4,...)의 효과를 통제한 후 확인하기 위해 다중회귀분석을 사용

- 용어
  - 단순회귀(simple regression or bivariate regression)
  - 다중회귀(multiple regression or multivariate regression)

다중회귀분석

두 개의 독립변수를 갖는 회귀식:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i, \quad i = 1,\ldots,n$$

- Y : 종속변수
- X_1, X_2 : 독립변수들
- β_0 = 모집단 수준의 절편
- β_1 = 회귀계수 1 (의 변화가 Y에 주는 효과, 는 고정!)
- β_2 = 회귀계수 2 (의 변화가 Y에 주는 효과, 는 고정!
- ε_i = 오차(누락된 변수들의 효과)

단순/다중회귀분석

예시)

- 단순회귀분석
  - $Wages = \beta_0 + \beta_1 * Adult\_Height + \varepsilon$
- 다중회귀분석
  - $Wages = \beta_0 + \beta_1 * Adult\_Height + \beta_2 * Adolescent\_Height + \tau$
  - $Wages = \beta_0 + \beta_1 * Adult\_Height + \beta_2 * Adolescent\_Height + \beta_3 * Athletics + \beta_4 * Club + u$

로지스틱 회귀분석

- 독립 변수의 선형 결합을 이용하여 사건의 발생 가능성을 예측하는데 사용되는 통계 기법

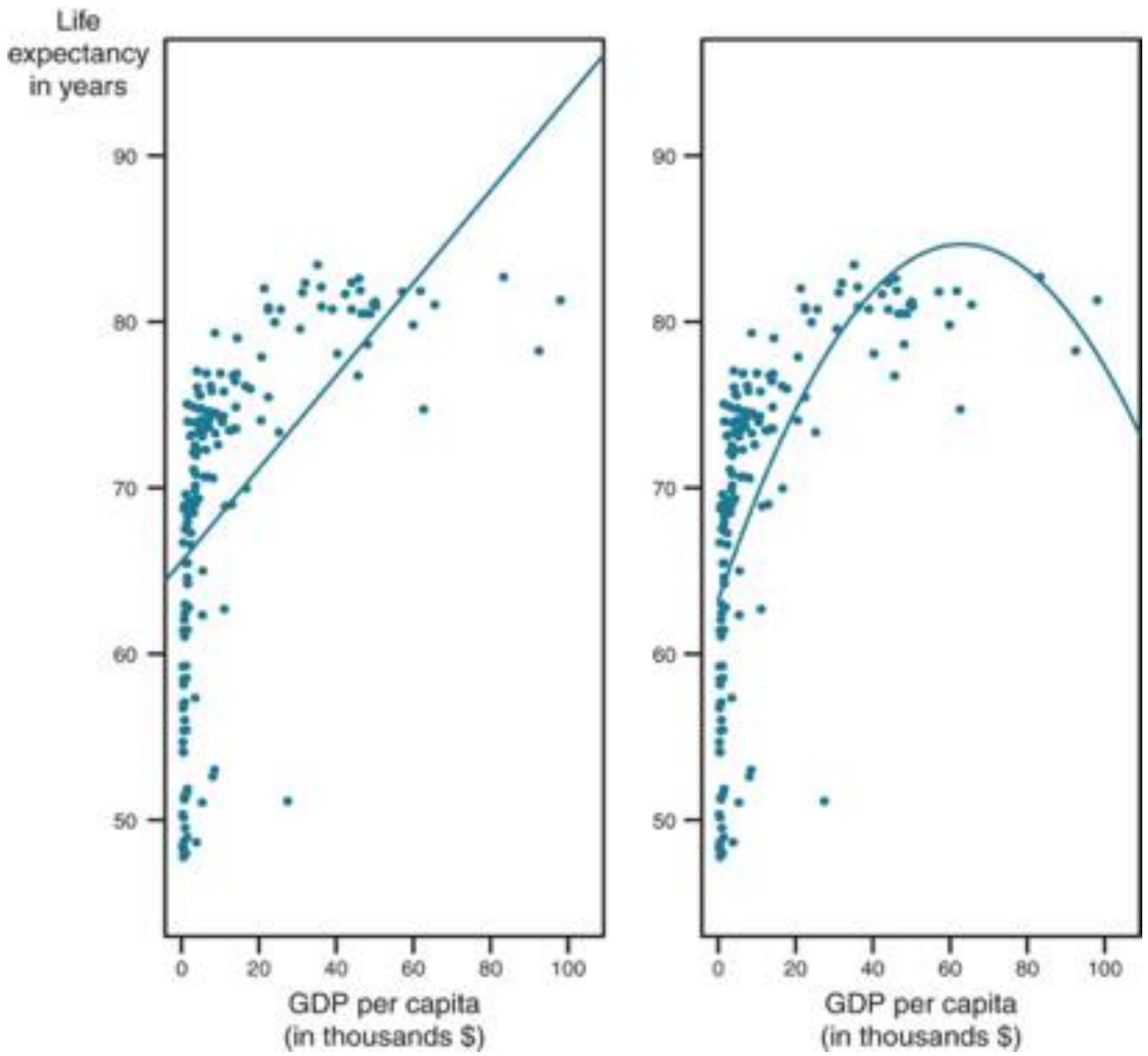
- 선형 회귀와의 **<u>유사점</u>**

-> 종속 변수와 독립 변수간의 관계를 구체적인 함수로 나타내어 향후 예측 모델에 사용

- 선형 회귀와의 **<u>차이점</u>**

-> 종속 변수가 범주형 데이터를 대상으로 하며 입력 데이터가 주어졌을 때 해당 데이터의 결과가 특정 분류로 나뉘기 때문에 일종의 분류 (classification) 기법

선형 모형(Linear Model)?



선형 모형(Linear Model)

- Ordinary Least Squre(OLS) Regression = Linear Model

- 여기서 선형이라 함은 회귀계수가 선형이라는 의미

- 회귀식에 포함되어 있는 독립변수, 종속변수는 반드시 선형일 이유가 없음

  - $Y = \beta_0 + (\beta_1)^2 X + \varepsilon$ (성립하지 않음)
  - $Y = \beta_0 + \beta_1 X_1 + \beta_2 (X_1)^2 + \varepsilon$ (성립함)

HYUNDAI MOBIS

선형 모형(Linear Model)

- $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$
- $y = \beta_0 x^{\beta_1}$
- $y = \frac{e^{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}}{1 + e^{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}}$

- $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \Rightarrow \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
- $y = \beta_0 x^{\beta_1} \Rightarrow log(y) = log(\beta_0 x^{\beta_1}) \Rightarrow log\beta_0 + \beta_1 \log(x) \Rightarrow y^* = \beta_0^* + \beta_1 x^*$
- $y = \frac{e^{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}}{1 + e^{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}} \Rightarrow \frac{y}{1-y} = e^{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3} \Rightarrow \log\left(\frac{y}{1-y}\right) = y^* = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$