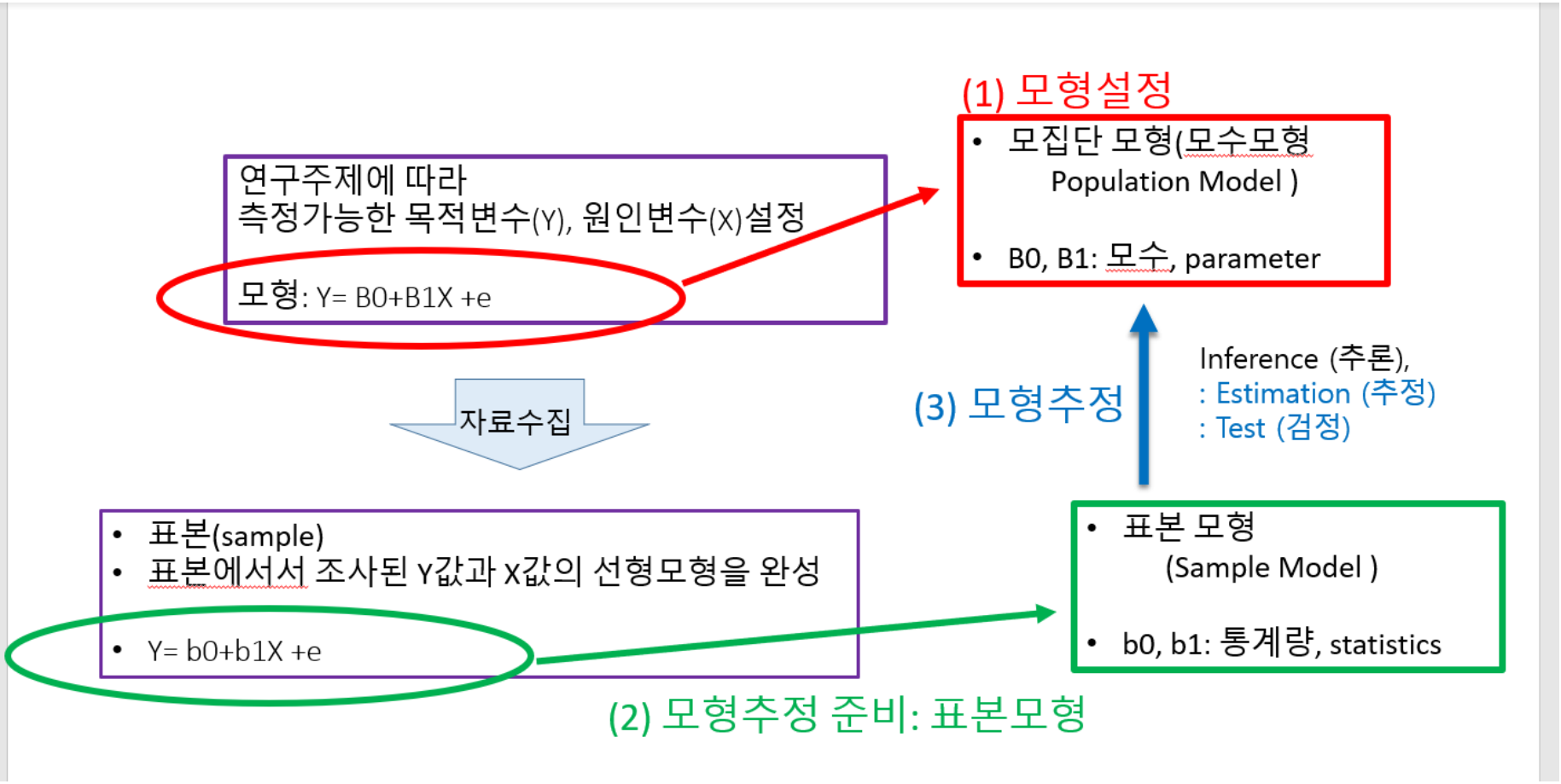
# Finance Analytics
# Chapter 3. Linear Regression Model
# Part 2. Model and Estimation

권태연

한국외대 국제금융학과

# List


(1) 모형설정
• 모집단 모형(모수모형
Population Model )
• B0, B1: 모수, parameter
연구주제에 따라
측정가능한 목적변수(Y), 원인변수(X)설정
모형: Y= B0+B1X +e
자료수집
(3) 모형추정
Inference (추론),
: Estimation (추정)
: Test (검정)
• 표본(sample)
• 표본에서서 조사된 Y값과 X값의 선형모형을 완성
• Y= b0+b1X +e
• 표본 모형
(Sample Model )
• b0, b1: 통계량, statistics
(2) 모형추정 준비: 표본모형

# 1. 모형설정(모형설계)

# 선형회귀모형 (Linear Regression Model)

$$Y_i = B_0 + B_1 X_{i1} + B_2 X_{i2} + ... + B_k X_{ik} + \epsilon_i$$

이때,

- $Y$는 종속변수(dependent variable), 반응변수(response variable), 피회귀자(regressand)
- $X_1, ...., X_2$들은 독립변수(independent variable), 설명변수 (explanatory variable, regressor), 공변량(covariate)
- $\epsilon$는 오차항 (error)
- 아랫첨자 $i$는 $i$번째 개체(individual) 의미함

⋆ X변수가 하나만 있는 모형을 단순선형회귀모형(Simple Linear Regression Model)

⋆ X변수가 두개 이상 있는 모형을 다중선형회귀모형(Multiple Linear Regression Model)

⋆ 단 Y변수는 언제나 한개만..

# 모집단 모형 (Population Model, True Model)

$$Y_i = B_0 + B_1 X_{i1} + B_2 X_{i2} + ... + B_k X_{ik} + \epsilon_i$$

Target variable $Y_i$에 대하여 다음의 두 가지 요소로 나누어 설명한다.

- 결정적 요소(deterministic), 체계적 요소 :
  $B_0 + B_1 X_{i1} + B_2 X_{i2} + ... + B_k X_{ik}$
- 무작위적 요소, 비체계적 요소 $\epsilon_i$

이때, 체계적 요소를

- $XB$ ( or 행렬을 이용한 표기방법:Y=XB+e)로 표기하기도 한다.
- $X_i$들 값이 주어져 있을때 $Y_i$의 평균적인 값이라는 의미에서 $Y_i$의 조건부 평균, 즉 $E(Y_i|X)$로 표기하기도 한다.
- (예)GDP와 기대수명:
  $Y_i = 68.716 + 0.420 x_i + e$
  $GDP =$ \$2만인 국가의 평균적 기대수명은?

# 회귀계수(Regression Coefficients)

모집단 모형

$$Y_i = B_0 + B_1 X_{i1} + B_2 X_{i2} + ... + B_k X_{ik} + \epsilon_i$$

에서 $B_0, B_1, ...B_k$를 **(모)회귀계수(regression coefficients),** 회귀모수 **(regression parameters)**라 한다.

- $B_0$은 상수항, 절편 (intercept)
- $B_1, ..B_k$는 기울기(slope coefficient)
- $B_0$의 의미?
- $B_1$의 의미?

# Slope Coefficient의 의미

- 설명변수 $X_1$값이 한 단위 변화하였을때, 반응변수 Y값이 $B_1$만큼 변화한다.
  - 단순 선형회귀모형에서: 위와 같이 해석
  - 다중 선형회귀모형에서: 다른 X변수들 $X_2, X_3, ..., X_k$값들이 일정할때라는 조건이후에 위와 같이 해석 = 편미분개념

# 2. 모형 추정 준비
# : 표본에서의 선형회귀모형

# Sample 모형

The sample counterpart is:

$$Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \ldots + b_k X_{ik} + e_i$$

Or, as written in short form:

$$Y = Xb + e$$

where $e$ is a residual(잔차). The deterministic component is written as

$$\hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \ldots + b_k X_{ik}$$

# 모집단 모형의 추정

모집단 모형

$$Y_i = B_0 + B_1 X_{i1} + B_2 X_{i2} + ... + B_k X_{ik} + \epsilon_i$$

을 표본 모형

$$Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \ldots + b_k X_{ik} + e_i$$

을 이용하여 추정, 추론한다.

- $B_0, ..., B_k$ : 모수
- $b_0, ..., b_k$ : 통계량, 추정량
- $B_0, ..., B_k$: 확률변수? 상수?
- $b_0, ..., b_k$: 확률변수? 상수?

# 변수 X와 오차항의 특성

- 설명변수 X는 확률변수가 아닌 상수로 가정한다.
- X변수를 확률변수로 정의하면 - 확률적 설명변수 (Chapter 19, Gujarati )
- 선형회귀모형에서 X변수와 Y변수와의 관계는 선형임을 가정한다.
- 정량변수, 범주형변수 모두 가능
- 일단.. Chapater 3를 배우는 동안은 X변수는 정량변수만을 고려한다.
- X변수로 범주형 변수 고려 - 질적 설명변수를 포함한 회귀모형 (Chpater 4. )
- 종속변수 Y는 확률변수.

⋆ 오차항의 특성

- 확률변수
- Y변수와 동일한 특성을 갖는다.

# 자료의 특성

1. 시계열자료 (Time Series Data)
   : 각기 다른 시점에서 관측된는 자료, 일별, 주별, 월별, 분기별, 연별.. 자료
   A set of observations that a variable takes at different times, such as daily (e.g., stock prices), weekly (e.g., money supply), monthly (e.g., the unemployment rate), quarterly (e.g., GDP), annually (e.g., government budgets), quinquenially or every five years (e.g., the census of manufactures), or decennially or every ten years (e.g., the census of population).
2. 횡단면자료 (Cross-Section Data)
   : 동시간대에 수집되는 하나또는 그 이상의 변수 자료
   Data on one or more variables collected at the same point in time. Examples are the census of population conducted by the Census Bureau every 10 years, opinion polls conducted by various polling organizations, and temperature at a given time in several places.

# 자료의 특성

3. 패널, 경시적 자료 (Panel, Longitudinal or Micro-panel Data) 시계열 + 횡단면 자료
   : Combines features of both cross-section and time series data. Same cross-sectional units are followed over time. Panel data represents a special type of pooled data (simply time series, cross-sectional, where the same cross-sectional units are not necessarily followed over time).

# 3. 선형회귀모형의 추정 (Estimation)

# 모형추정

모집단 모형

$$Y_i = B_0 + B_1 X_{i1} + B_2 X_{i2} + ... + B_k X_{ik} + \epsilon_i$$

을 표본 모형

$$Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \ldots + b_k X_{ik} + e_i$$

을 이용하여 추정, 추론한다.

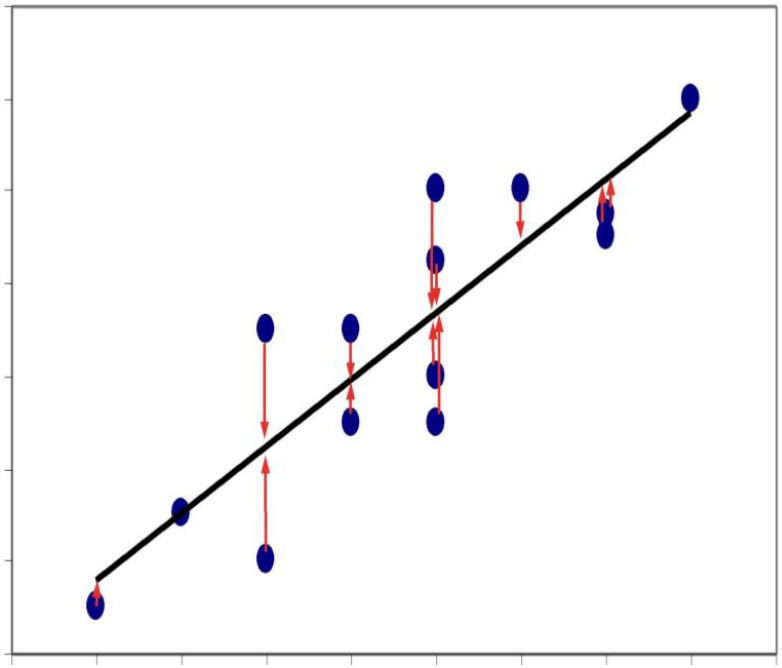
$$b_0, b_1, b_2, ..., b_k \rightarrow B_0, B_1, B_2, ...B_k$$

How? 비체계적 부분의 최소화... (why? )

# 단순 선형회귀모형에서 직관적으로 추정방법이해하기

$$b_0, b_1 \rightarrow B_0, B_1$$

How? 비체계적 부분의 최소화



# 최소제곱추정법 (Least Square Estimation)

- (통상) 최소제곱추정법 ( (Ordinary) Least Square Estimation) 방법은 선형회귀모형의 비체계적 부분인 오차 $u_i$들을 최소화하는 $B_0, B_1, B_2, ...B_k$에 대한 추정량 $b_0, b_1, b_2, ..., b_k$을 구하는 방법이다. 형
- y로부터 $\hat{y}$의 거리 $u_i$의 제곱의 합을 최소화하는 추정량 $b_0, b_1, b_2, ..., b_k$을 구하는 방법
- 간략히 OLS라 한다.

# 단순선형회귀모형에서의 OLS

다음과 같은 단순선형회귀모형 (모집단 모형)이 있다고 하자.

$$Y_i = B_0 + B_1 X_{i1} + u_i$$

다음과 같은 오차항의 제곱합 (error sum of square, ESS) or (sum of squared error, SSE)

$$\sum_{i=1}^{i=n} u_i^2 = \sum_{i=1}^{i=n} (Y_i - B_0 - B_1 X_{i1})^2$$

를 최소화하는 $b_0$, $b_1$값을 $B_0$, $B_1$값의 최소제곱 추정량(OLS)라 한다.
ESS(or SSE)를 최소화 하는 $b_0$, $b_1$을 구하는 방법?
: 미분
: 편미분

- 1차 편미분=0
- 2차 편미분 $> 0$

# 단순선형회귀모형에서의 OLS estimators

$$b_1 = r\frac{s_y}{s_x} \quad (1)$$

$$b_0 = \bar{y} - b_1\bar{x} \quad (2)$$

# 다중형선형회귀모형에서의 OLS

다음과 같은 다중 선형회귀모형 (모집단 모형)이 있다고 하자.

$$Y_i = B_0 + B_1 X_{i1} + B_2 X_{i2} + ... + B_k X_{ik} + \epsilon_i$$

다음과 같은 오차항의 제곱합 (error sum of square, ESS) or (sum of squared error, SSE)

$$\sum_{i=1}^{i=n} \epsilon_i^2 = \sum_{i=1}^{i=n} (Y_i - B_0 - B_1 X_{i1} - ... - B_k X_{ik})^2$$

를 최소화하는 $b_0$,... ,$b_k$값을 $B_0$, $B_k$값의 최소제곱 추정량(OLS)라 한다.

행렬로 접근...

# 다중 선형회귀모형에서의 OLS estimators

$$b = (b_0, b_1, ...b_k)' = (X'X)^{-1}X'Y \tag{3}$$

# Summary:

⋆ 모형설계: Population regression model is:

$$Y_i = B_0 + B_1 X_{i1} + B_2 X_{i2} + ... + B_k X_{ik} + \epsilon_i = XB + \epsilon$$

이때,

- $Y$는 종속변수(dependent variable), $X_1, ...., X_2$들은 독립변수 (independent variable), 설명변수(explanatory variable, regressor), 공변량(covariate)
- $\epsilon$는 오차항 (error)

⋆ 모형추정: The sample counterpart is:

$$Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \ldots + b_k X_{ik} + e_i = Xb + e$$

이때,

- $e_i$ is a residual(잔차).
- $b_0$,..$b_k$는 회귀계수(regression coefficients) $B_0, ..B_k$의 추정량 (estimator)
- 자료를 수집하여 추정량에 숫자를 대입하여 구체적 값을 구하면 회귀계수의 추정치(estimate)
- 추정방법: OLS