Chapter 2　Concept of Regression

남준우·허인(2021), 제3장

Gujarati/Porter(2018), 제2장

(1) Relationship between economic variables

(2) Theoretical(Economic) model vs. Statistical(Econometric) model

(3) Population Regression Function(PRF)

(4) Stochastic Presentation of PRF

(5) Sample Regression Function(SRF), Sample Regression Line

(6) Least Squares Method

(7) Prediction, Forecasting

(1) Relationship between economic variables

- $\Delta X \Rightarrow \Delta y$

(Examples)

Δ income ⇒ Δ consumption

Δ money supply ⇒ Δ inflation

Δ advertisement ⇒ Δ sales

Δ 노선별 경쟁상태, 항공 거리, 지역 등 ⇒ Δ 항공권 가격

Δ 투수의 실적 ⇒ Δ 투수의 연봉

Δ 교육비지출, 여성의 취업률, 혼인율 등 ⇒ Δ 출산율

Δ 영화의 특성 ⇒ Δ 흥행(혹은 관객수)

Δ 국가별 특성 ⇒ Δ 자살률

Δ 도시별 특성 ⇒ Δ 빈곤율

- <u>Modeling economic behavior</u>: $y = f(X)$

► Simplification: $y = \beta_1 + \beta_2 X$

► $\beta_2 = \dfrac{dy}{dX}$: marginal effect of X on y

(2) Theoretical(Economic) model vs. Statistical(Econometric) model

- Theoretical model: model for specific individual

$$y_i = f(X_i)$$

► model for representative individual

- Statistical model: model for average or systematic behavior of many individuals or firms

$$y_i = f(X_i) + \varepsilon_i$$

(3) Population Regression Function(PRF)

① (univariate case) $\mu = E(y_i)$

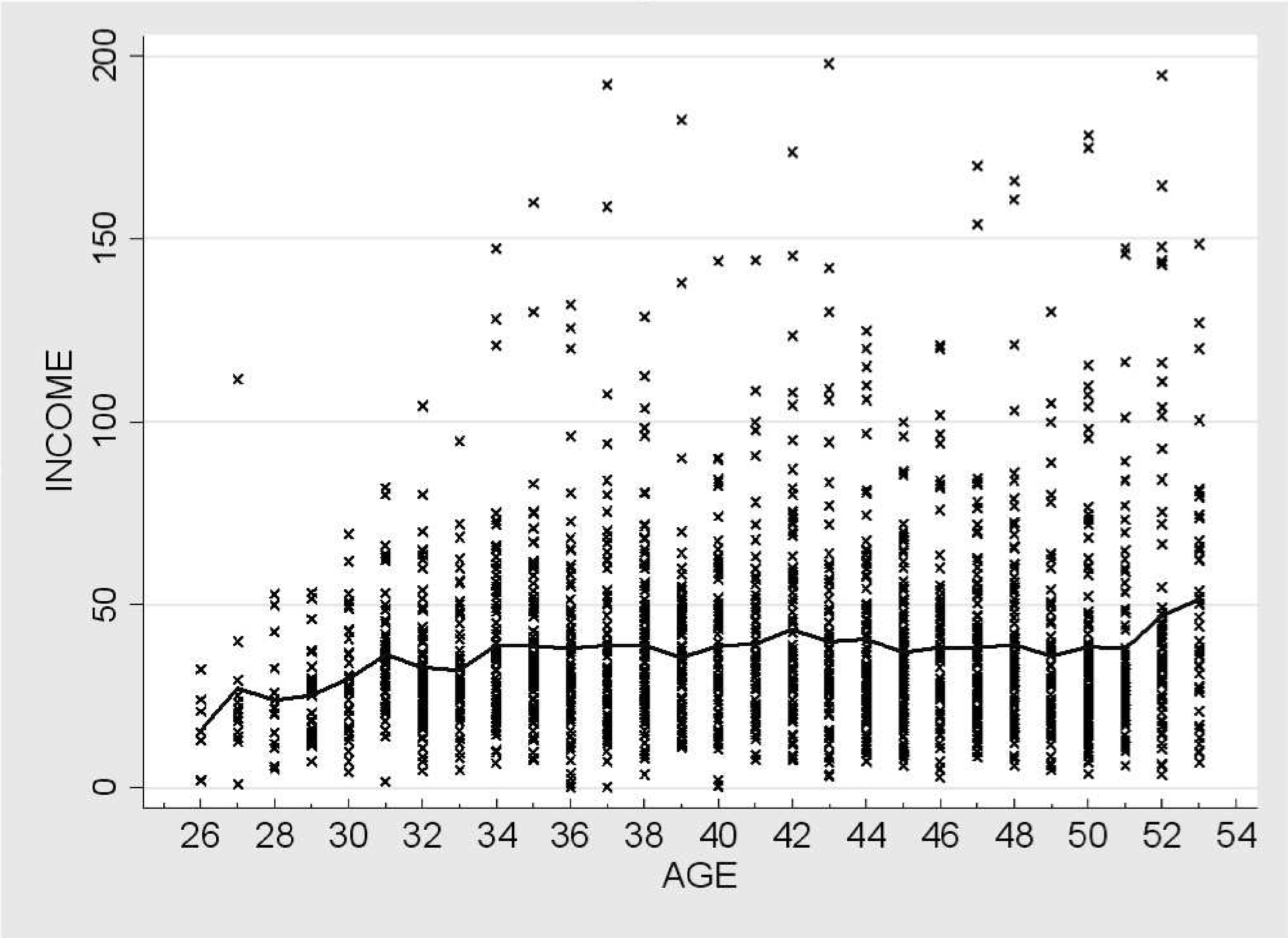
then $y_i = \mu + \varepsilon_i$

② (Bivariate example) income vs. age

► Conditional mean, conditional expectation

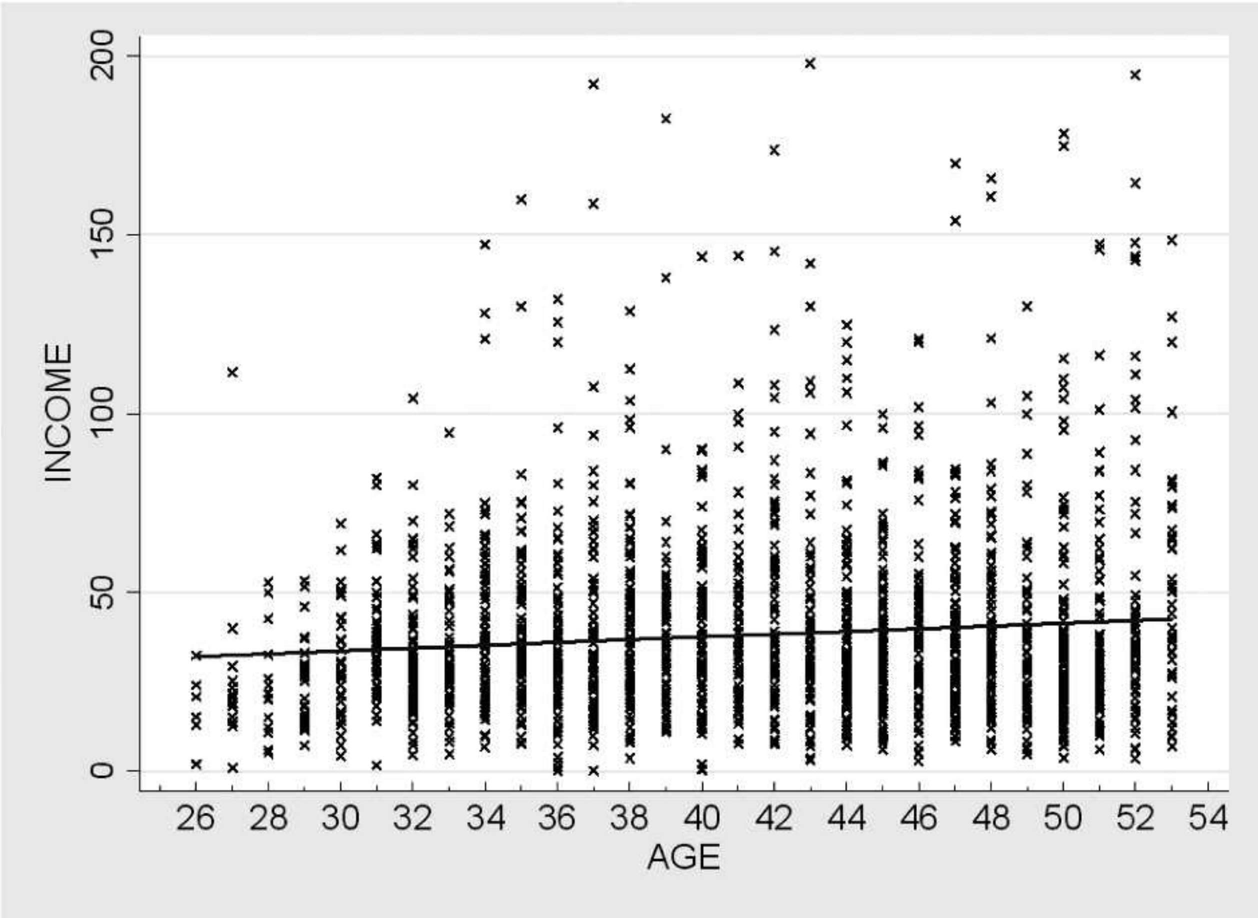
► Conditional expectation function

$$y_i = E(y_i | X_i) + \varepsilon_i$$



► Conditional Expectation Function(CEF), PRF, Population Regression Line

- Assume PRF is linear,



► $E(y_i | X_i) = \beta_1 + \beta_2 X_i$; Linear Population Regression Line

► (β_1, β_2): Regression coefficient

► β_1: intercept coefficient

► β_2: slope coefficient $\Rightarrow$ marginal effect of X on y

$\Rightarrow$ unknown, to be estimated

► The term of ‘regression’

(4) Stochastic Presentation of PRF

- Error term

  ► $E(y_i \mid X_i) = \beta_1 + \beta_2 X_i$

  ► $y_i = E(y_i \mid X_i) + \varepsilon_i$

$\Rightarrow \quad y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$

► y_i: dependent variable, endogenous

► X_i: independent variable, explanatory variable, exogenous

► (β_1, β_2): regression coefficient

► ε_i: error term

- $\varepsilon_i = y_i - E(y_i \mid X_i) = y_i - (\beta_1 + \beta_2 X_i)$

$\Rightarrow \; E(\varepsilon_i \mid X_i) = 0$

- <u>Sources of error term</u>

① Omitted variables

② Approximation error

③ Measurement error

④ Real unpredictable error

(5) Sample Regression Function(SRF), Sample Regression Line

; Estimation of Population Regression Function

- population data vs. sample data

① SRF, Sample Regression Line

- $y_i = b_1 + b_2 X_i + e_i$

- $\hat{y}_i$: fitted value
- (b_1, b_2): estimator of (β_1, β_2)
- e_i: residual

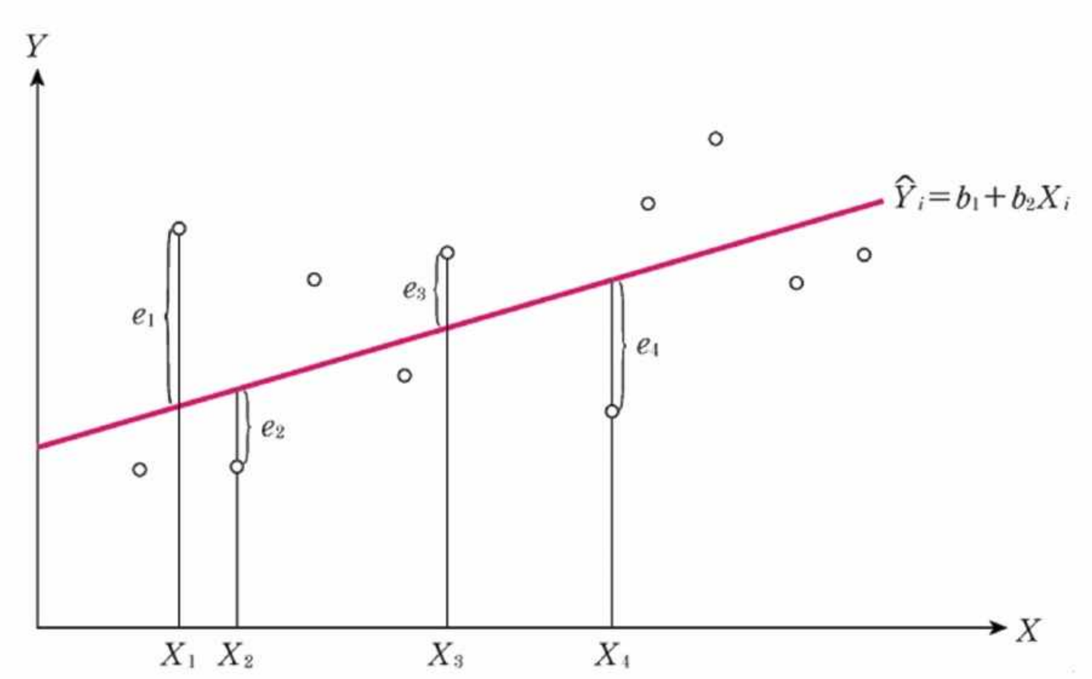
② $\beta_1 = b_1,\ \beta_2 = b_2$ can NOT be guaranteed.

③ (b_1, b_2) are random variables.

► We have to configure the sampling distribution of (b_1, b_2).

(6) Least Squares Method: How to get (b_1, b_2)?

- Choose (b_1, b_2) which minimizes $\sum_{i=1}^{n} e_i^{\,2} = \sum_{i=1}^{n} (y_i - b_1 - b_2 X_i)^2$: Least Squares Estimator



- Why Least Squares?

► Why not minimizing $\sum_{i=1}^{n} e_i$?

► What about minimizing $\sum_{i=1}^{n} |e_i|$? Least Absolute Deviations(LAD) Estimator

◎ Least Squares Estimator

Choose (b_1, b_2) which minimizes $\sum_{i=1}^{n} e_i^{\,2} = \sum_{i=1}^{n} (y_i - b_1 - b_2 X_i)^2$.

F.O.C.:

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} \left(Y_1 - b_1 - b_2 X_i \right) = 0$$

$$\sum_{i=1}^{n} X_i e_i = \sum_{i=1}^{n} X_i \left(Y_1 - b_1 - b_2 X_i \right) = 0$$

$$b_2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(y_i - \overline{y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} = \frac{\sum_{i=1}^{n} X_i y_i - n\overline{X}\,\overline{y}}{\sum_{i=1}^{n} X_i^{\,2} - n\overline{X}^{\,2}} = \frac{S_{Xy}}{S_X^{\ 2}}$$

$$b_1 = \overline{y} - b_2\overline{X}$$

(Example)

- y:소비, X: 소득.

obs	Y_i	X_i	$(X_i-\bar{X})^2$	$(X_i-\bar{X})(Y_i-\bar{Y})$	$\hat{Y}_i$	e_i	e_i^2	$(Y_i-\bar{Y})^2$	$(X_i-\bar{X})X_i$	$(X_i-\bar{X})Y_i$
1	70	80	8100	3690	65.28	4.82	23.21	1681		
2	65	100	4900	3220	75.36	-10.36	107.41	2116		
3	90	120	2500	1050	85.55	4.45	19.84	441		
4	95	140	900	480	95.73	-0.73	0.53	256		
5	110	160	100	10	105.91	4.09	16.74	1		
6	115	180	100	40	116.09	-1.09	1.19	16		
7	120	200	900	270	126.27	-6.27	39.35	81		
8	155	240	4900	3080	146.64	8.36	69.95	1936		
9	150	260	8100	3510	156.82	-6.82	46.49	1521		
10	140	220	2500	1450	136.45	3.55	12.57	841		
합	1110	1700	33000	16800		0	337.27	8890		

► $\bar{Y}=1110/10=111, \qquad \bar{X}=1700/10=170$

- $b_2=\dfrac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}=\dfrac{16800}{33000}=0.51,$
- $b_1=\bar{Y}-b_2\bar{X}=111-0.51\times 170=24.5$

(Remarks)

① LS estimator vs. LS estimate

② If X_i has only one variable ($X_i = c$ for all i), then b_2 can NOT be obtained.

► Identification condition

③ Regression line passes thru $(\bar{X}, \bar{y})$.

④ $S_{Xe} = 0$ and $S_{\hat{y}e} = 0$.

⑤ Linear regression: We require ‘linear in (β_1, β_2)’.

We do NOT require ‘linear in X’.

(Definition of linear) $f(x)$ is linear in x, if $f'(x)$ is not a function of x.

(Examples)

① $y_i = \beta_1 + \beta_2 \ln X_i + \varepsilon_i$; linear

② $y_i = \beta_1 + \sqrt{\beta_2} X_i + \varepsilon_i$; nonlinear

③ $y_i = \beta_1 + \beta_2 X_i^{\,2} + \varepsilon_i \, \beta_2$; linear

(7) Prediction, Forecasting

• Suppose, for some observation f, the value of X is known as X_f,

the best predictor of y_f is

$$\hat{y}_f = b_1 + b_2 X_f .$$

(Example) In consumption function example, $X_f = 200$,

then $\hat{y}_f = 24.5 + 0.51 \times 200 = 126.25$