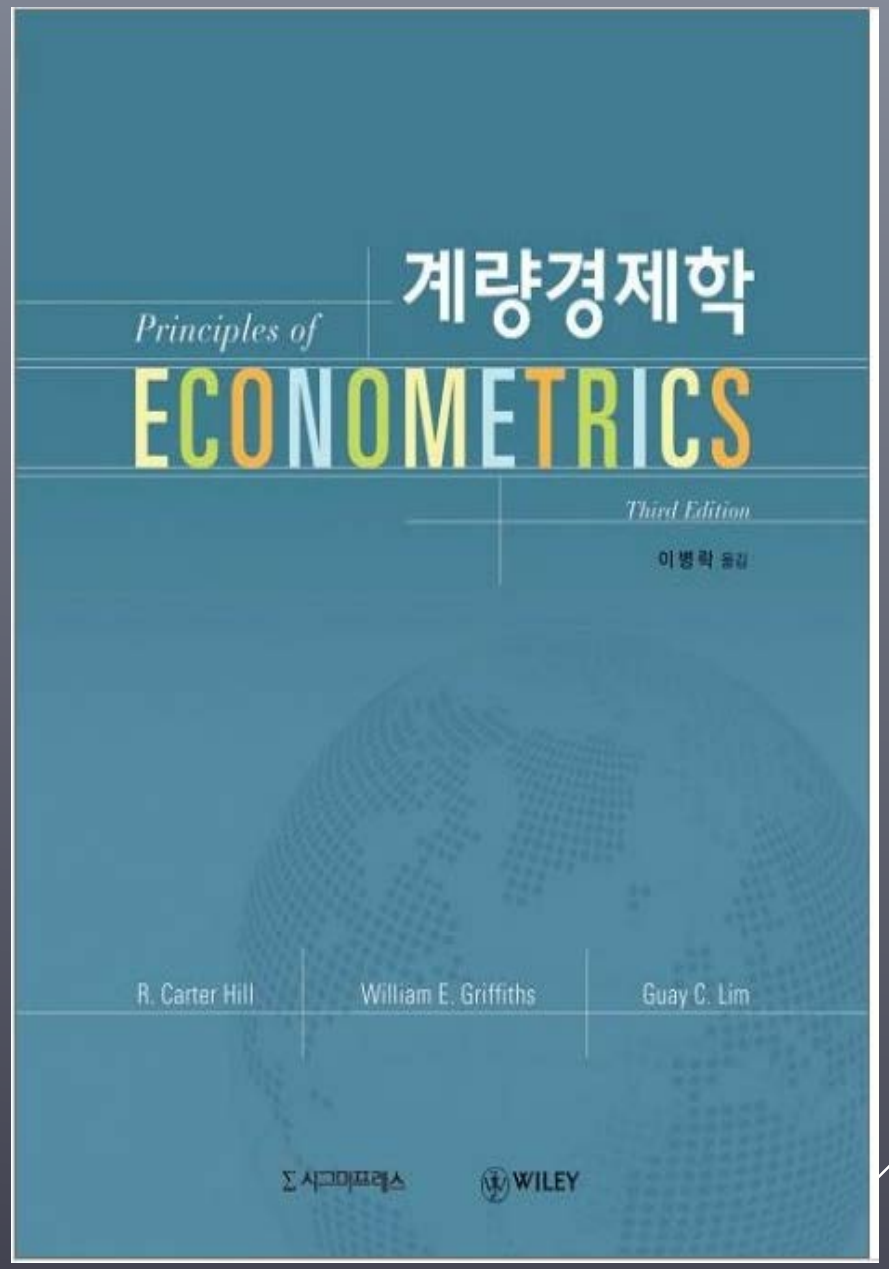


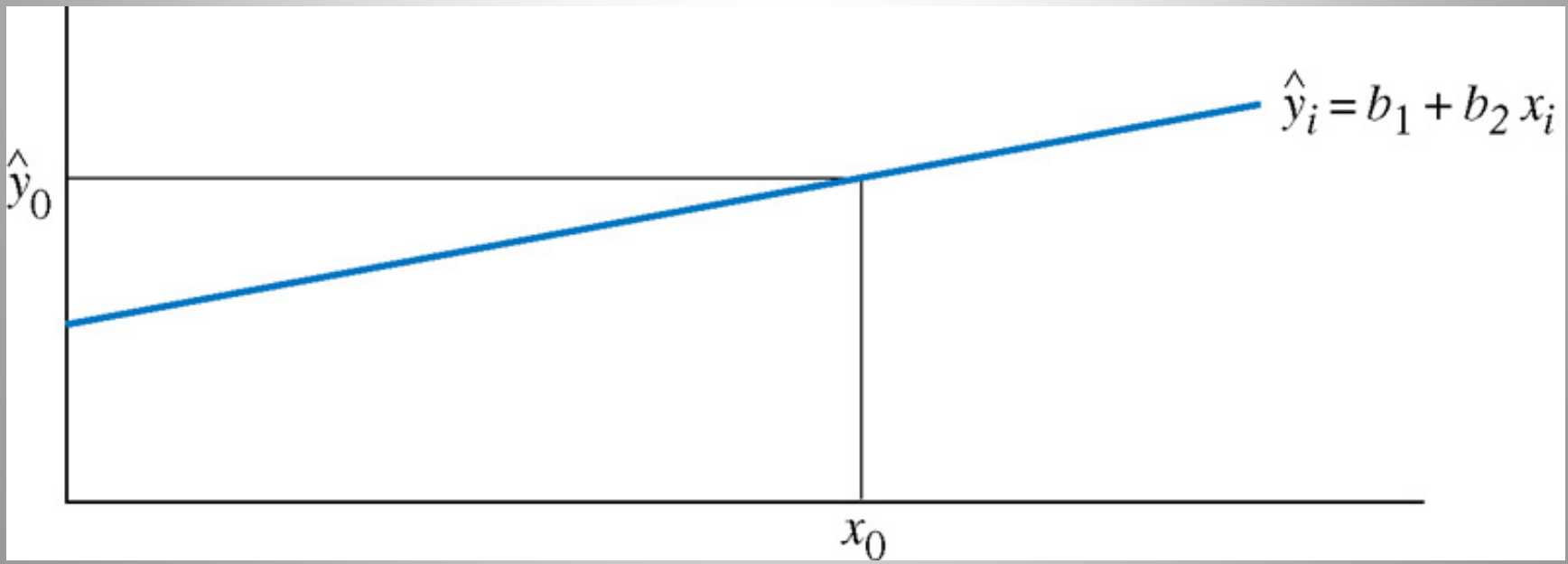
제4장
예측, 적합도, 모형화

4.1 OLS 예측

(1) 점예측

- 설명변수 x_0 일 때, 종속변수 y_0 의 값을 예측하고자 함

$$\hat{y}_0 = b_1 + b_2 x_0$$



- **예측오차(forecast error), f**

$$f = y_0 - \hat{y}_0 = (\beta_1 + \beta_2 x_0 + e_0) - (b_1 + b_2 x_0)$$

- 예측오차의 기대값

$$E(f) = \beta_1 + \beta_2 x_0 + E(e_0) - [E(b_1) + E(b_2) x_0]$$

$$= \beta_1 + \beta_2 x_0 + 0 - [\beta_1 + \beta_2 x_0] = 0$$

- 예측오차의 분산

$$\text{var}(f) = \sigma^2 \left[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]$$

- $\hat{y}_0$ 는 y_0 의 최우수 선형 불편 예측값(BLUP)

Best
Linear
Unbiased
Predictor

- **예측오차(forecast error)의 분산을 줄이기 위한 방법**

= 예측의 정확도를 높이기 위한 방법

$$\text{var}(f) = \sigma^2 \left[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]$$

- 모형의 전체적 불확실성(σ^2)을 줄임
- 표본의 크기(N)를 늘림
- 설명변수의 변동을 크게 함
- 평균 근처에서 예측함

  (평균에서 벗어나서 예측할수록 정확도가 떨어짐)

(2) 구간예측

$$\hat{y}_0 \pm t_c \mathrm{se}(f)$$

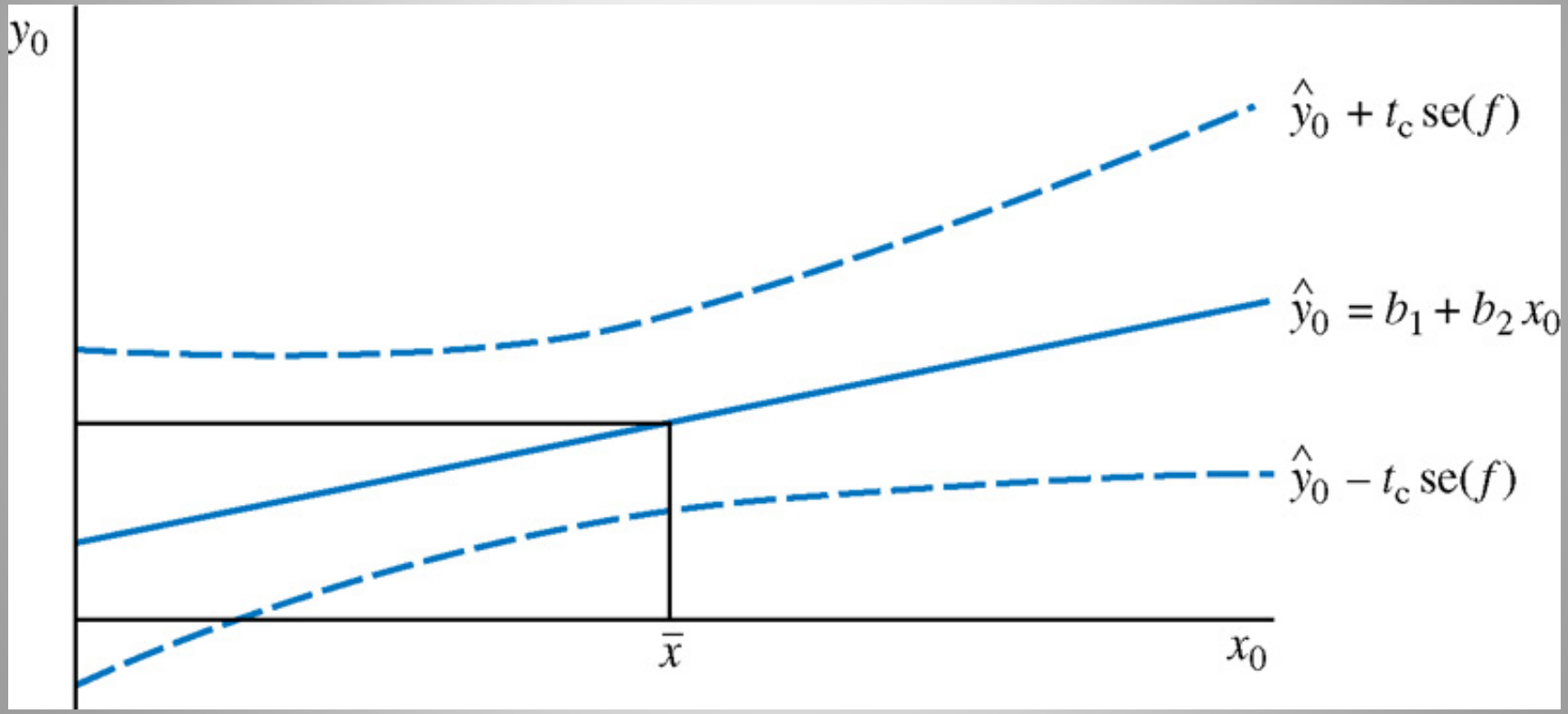
- 신뢰구간, 혹은 예측구간

$$P[\hat{y}_0 - t_c \mathrm{se}(f) \le y_0 \le \hat{y}_0 + t_c \mathrm{se}(f)] = 1 - \alpha$$

$$\mathrm{se}(f) = \sqrt{\widehat{\mathrm{var}(f)}} \quad \widehat{\mathrm{var}(f)} = \hat{\sigma}^2 \left[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]$$

▪ 구간예측

- x_0가 $\bar{x}$ 에서 멀어질수록, 예측오차 분산 증가 ⇒ 예측 신뢰성 감소



4.1.1 식료품 지출액 모형의 예측

✓ 주당 소득이 $2,000인 가계의 식료품 지출액은 얼마인가?

- 점예측 $\hat{y}_0 = b_1 + b_2 x_0 = 83.4160 + 10.2096(20) = 287.6089$
- 구간예측

$$\hat{y}_0 \pm t_c \text{se}(f) = 287.6069 \pm 2.0244(90.6328) = [104.1323, 471.0854]$$

$$\widehat{\text{var}(f)} = \hat{\sigma}^2 \left[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]$$

$$= \hat{\sigma}^2 + \frac{\hat{\sigma}^2}{N} + (x_0 - \bar{x})^2 \frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}$$

$$= \hat{\sigma}^2 + \frac{\hat{\sigma}^2}{N} + (x_0 - \bar{x})^2 \widehat{\text{var}(b_2)}$$

4.2 적합도의 측정

$$Y_t = \beta_1 + \beta_2 X_t + e_t$$

- 이 모형을 추정하고 분석하는 이유의 하나는 '예측'임
- 설명변수 X_t가 모형에 도입된 이유는
  그것이 종속변수 Y_t 의 변동을 가능한 많이 설명할
  것이라고 생각하기 때문임

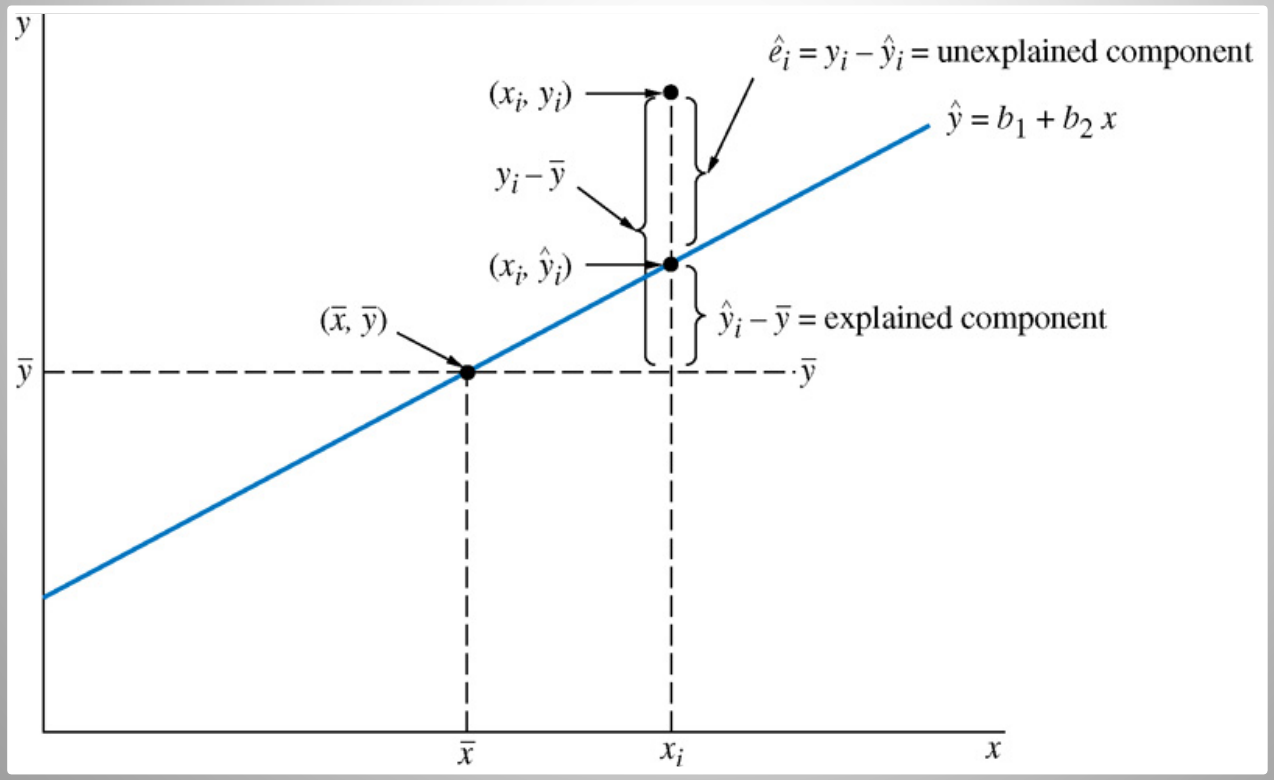
- $E(Y_t) = \beta_1 + \beta_2 X_t$: 체계적인 요소, 설명할 수 있는 부분
- e_t : 무작위적 교란요소, 설명할 수 없는 부분

✓ 이 두 요소는 실제로는 관찰 불가능

- 추정을 통하여 관찰 가능한 식

$$Y_t = \hat{Y}_t + \hat{e}_t \qquad \text{여기서} \quad \hat{Y}_t = b_1 + b_2 X_t, \quad \hat{e}_t = Y_t - \hat{Y}_t$$

- Y_t 의 총변동을 측정하는 합리적인 방법: $\sum (Y_t - \bar{Y})^2$



p.131

<그림 4.3>

$$\sum (y_t - \bar{y})^2 = \sum [(\hat{y}_t - \bar{y}) + \hat{e}_t]^2$$

$$= \sum (\hat{y}_t - \bar{y})^2 + \sum \hat{e}_t^2 + 2\sum (\hat{y}_t - \bar{y})\hat{e}_t$$

$$= \sum (\hat{y}_t - \bar{y})^2 + \sum \hat{e}_t^2$$

- The cross-product term $\sum (\hat{y}_t - \bar{y})\hat{e}_t = 0$ and drops out.

1. $\sum (y_t - \bar{y})^2$ = total sum of squares = *SST*: a measure of *total variation* in y about its sample mean.
2. $\sum (\hat{y}_t - \bar{y})^2$ = explained sum of squares = *SSR*: that part of total variation in y about its sample mean that is explained by the regression.
3. $\sum \hat{e}_t^2$ = error sum of squares = *SSE*: that part of total variation in y about its mean that is not explained by the regression.

$$\sum (Y_t - \bar{Y})^2 = \sum (\hat{Y}_t - \bar{Y})^2 + \sum \hat{e}_t^2$$

$$SST = SSR + SSE$$

- 총변동을 이렇게 분해하여 정리한 것을 **분산분석표**(analysis of variance table)라 함

Analysis of Variance Table

Source of Variation	DF	Sum of Squares	Mean Square
Explained	1	*SSR*	*SSR*/1
Unexplained	*T*–2	*SSE*	*SSE*/(*T*–2) $[=\hat{\sigma}^2]$
Total	*T*–1	*SST*	

- **결정계수** (coefficient of determination)

  : 종속변수의 변동 중 설명변수로 회귀모형 내에서 설명할 수 있는 부분의 비율

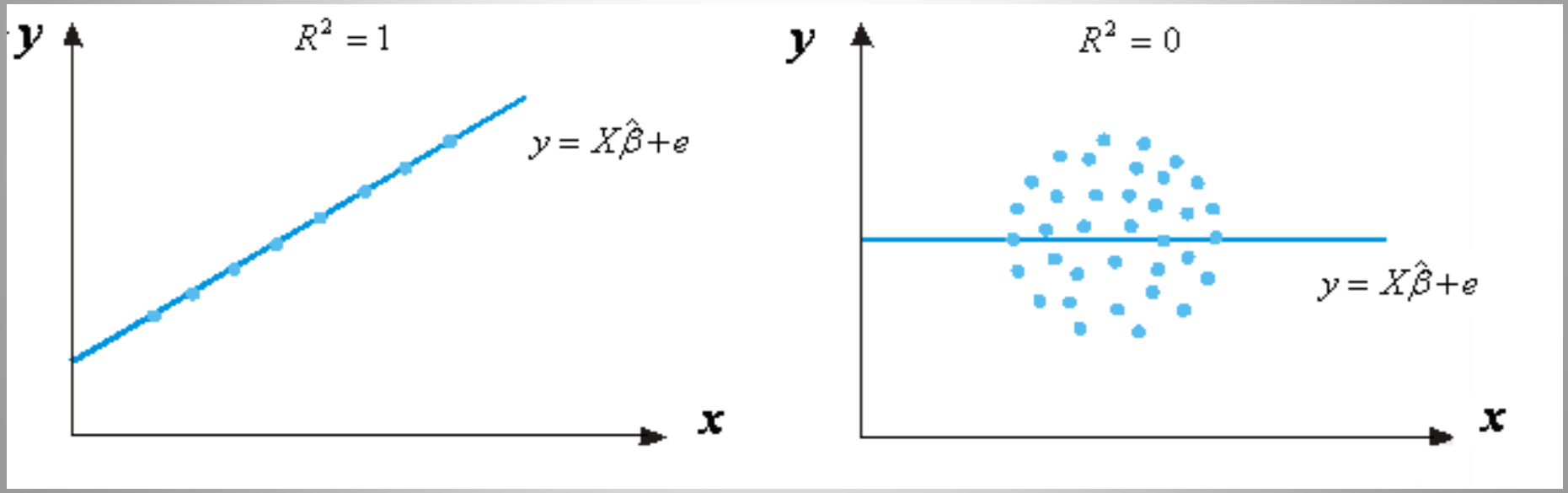
$$SST = SSR + SSE$$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- R^2 가 1에 가까울수록 $\hat{Y}_t = b_1 + b_2 X_t$ 로 Y_t 의 변동을 잘 설명할 수 있다는 것을 의미함,
  또 추정된 모형의 예측 성과(능력)가 높다는 것을 의미함

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$0 \le R^2 \le 1$$



- **유의사항**

- R^2 만으로는 회귀모형의 질을 측정할 수 없음
- R^2 를 극대화하는 것에만 초점을 맞추는 것은 올바른 회귀분석 방법이 아님
- R^2 는 단지 회귀모형의 **적합도(goodness of fit)**를 측정하여 나타내는 척도임

4.2.3 식료품 지출액 추정식의 적합도 측정

$$SST = \sum \left(y_i - \bar{y} \right)^2 = 495132.160$$

$$SSE = \sum \left(y_i - \hat{y}_i \right)^2 = \sum \hat{e}_i^2 = 304505.176$$

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{304505.176}{495132.160} = 0.385$$

- 해석
- 식료품 지출액 변동은 약 39%가 소득 변동으로 설명 가능
- 추정된 회귀모형은 식료품 지출액 변동의 약 39%를 설명할 수 있으나, 변동의 61%는 이 모형으로 설명할 수 없다는 의미

<EViews Regression Output>

Dependent Variable: *FOOD_EXP*
Method: Least Squares
Sample: 1 40
Included observations: 40

	Coefficient	Std. Error	*t*-Statistic	Prob.
C	83.41600	43.41016	1.921578	0.0622
INCOME	10.20964	2.093264	4.877381	0.0000

R-squared	0.385002	Mean dependent var	283.5735
Adjusted R-squared	0.368818	S.D. dependent var	112.6752
S.E. of regression	89.51700	Akaike info criterion	11.87544
Sum squared resid	304505.2	Schwarz criterion	11.95988
Log likelihood	−235.5088	Hannan-Quinn criter	11.90597
F-statistic	23.78884	Durbin-Watson stat	1.893880
Prob(F-statistic)	0.000019		

<Excel Regression Output>

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.620485472							
5	R Square	0.385002221							
6	Adjusted R Square	0.368818069							
7	Standard Error	89.51700429							
8	Observations	40							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	190626.9788	190626.9788	23.78884107	1.94586E-05			
13	Residual	38	304505.1742	8013.294058					
14	Total	39	495132.153						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	83.41600997	43.41016192	1.921577951	0.062182379	-4.463267721	171.2952877	-4.463267721	171.2952877
18	income	10.2096425	2.093263461	4.877380554	1.94586E-05	5.972052202	14.4472328	5.972052202	14.4472328
19									
20									
21									
22									
23									
24									
25									
26									
27									
28									
29									
30									
31									
32									

food exp plot | **Food Expenditure** | Sheet1

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{304505.176}{495132.160} = 0.385$$

- **유의사항**

$$R^2 = 0.385$$

This R^2 value may seem low but it is typical in studies involving **cross-sectional** data analyzed at the individual or **micro** level.

A considerably higher R^2 value would be expected in studies involving **time-series** data analyzed at an aggregate or **macro** level.

4.2.2 상관계수와 결정계수의 관계

- correlation coefficient ρ between *X and Y*

$$\rho = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\,\mathrm{var}(Y)}}$$

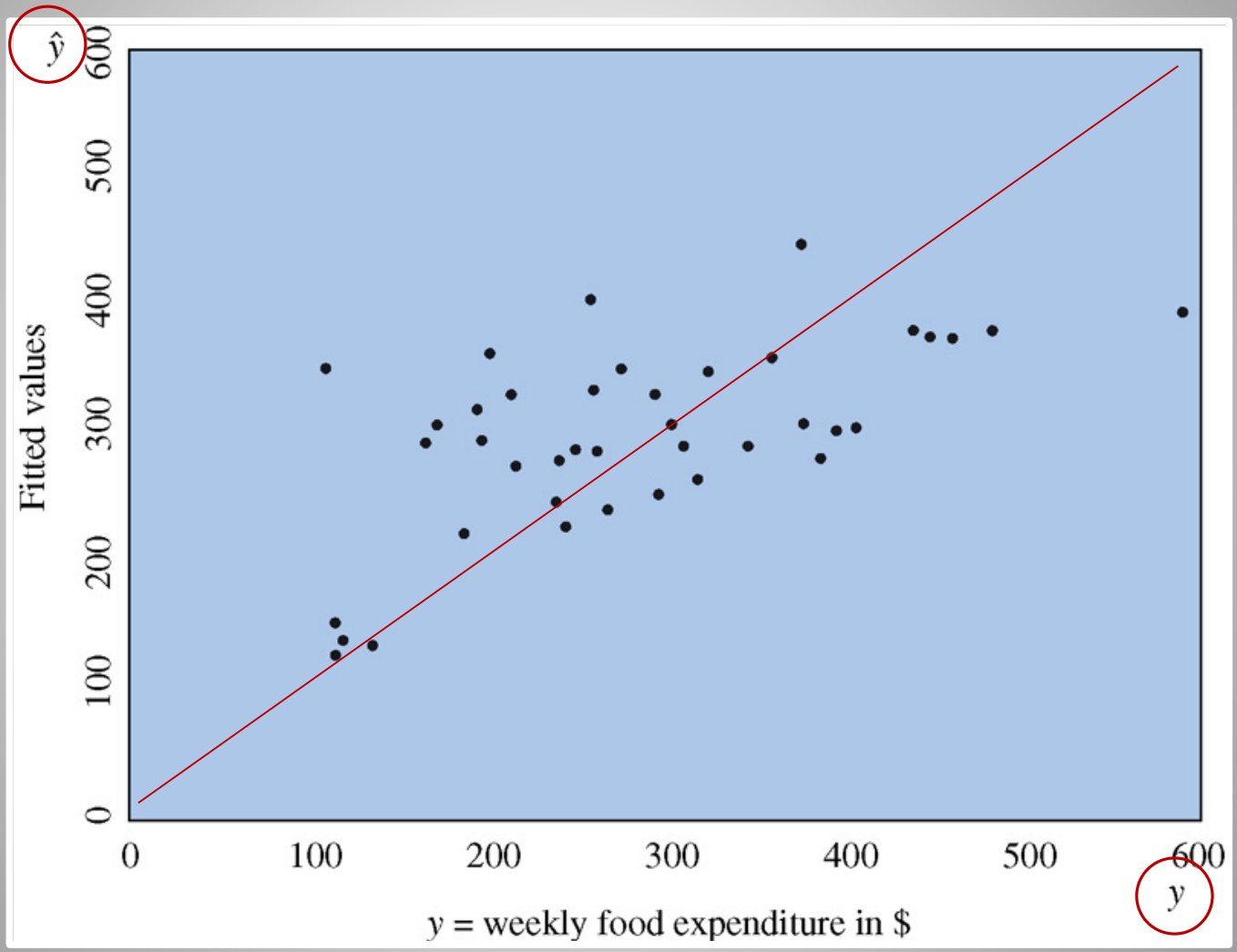
- sample correlation coefficient

$$r = \frac{\hat{\mathrm{cov}}(X,Y)}{\sqrt{\hat{\mathrm{var}}(X)\,\hat{\mathrm{var}}(Y)}}$$

$$r = \frac{\sum_{t=1}^{T}(x_t-\bar{x})(y_t-\bar{y})}{\sqrt{\sum_{t=1}^{T}(x_t-\bar{x})^2\sum_{t=1}^{T}(y_t-\bar{y})^2}}$$

- $r_{XY}^2 = R^2$
- $r_{Y\hat{Y}}^2 = R^2$: R^2 is called a measure of "goodness of fit."

- $r^2_{Y\hat{Y}} = R^2 = 0.385$



4.2.4 회귀분석 결과를 보고하는 방법

- 회귀분석 컴퓨터 SW는 다양한 정보를 제공함
- 회귀분석 결과를 보고할 때는 그러한 정보를 아래와 같이 요약하여 표시함

$$FOOD\_EXP = 83.42 + 10.21\, INCOME \qquad R^2 = 0.385$$
$$\text{(se)} \qquad (43.41)^{*} \quad (2.09)^{***}$$

$$FOOD\_EXP = 83.42 + 10.21\, INCOME \qquad R^2 = 0.385$$
$$\text{(t-값)} \qquad (1.92)^{*} \quad (4.88)^{***}$$

- *, **, *** indicates significant at the 10%, 5%, 1% level, respectively.

4.3 모형화 문제

- **통계자료를 일정한 비율로 조절하면?**

- 통계수치가 너무 큰 경우는 편리한 측정단위로 조절하여 발표함

(예) $X = 93{,}491{,}400{,}000{,}000\,(\text{원}) \Rightarrow X^* = 93.4914\,(\text{조원})$

이 경우 자료를 다음과 같은 비율로 조절한 것임

$$X^* = X/1{,}000{,}000{,}000{,}000$$

- 자료를 일정한 비율로 조절하는 경우,

⇒ 추정결과에 어떤 영향을 미치는가?

⇒ 추정결과의 해석에는 어떤 영향을 미치는가?

▪ X 의 측정단위 변화 경우

(예) $X_t^* = X_t/100$ 로 조절한 경우

$$\begin{aligned}\hat{y}_t &= 40.77 + 0.1283 x_t \\ &= 40.77 + (100 \times 0.1283)\left(\frac{x_t}{100}\right) \\ &= 40.77 + 12.83 x_t^*\end{aligned}$$

- 소득이 \$100 증가하면, 식료품 지출은 \$12.83 증가함

$$if\ \Delta X = 100 \ \text{즉}\ \ \Delta X^* = 1, \quad then \quad \Delta Y = 12.83.$$

- 해당 회귀계수 및 표준오차(b_2, $se(b_2)$)도 100배 증가
- 다른 모든 회귀분석 통계량은 불변

▪ Y 의 측정단위 변화 경우

(예) $Y_t^* = 100 \times Y_t$ 로 조절한 경우

$$100\hat{Y}_t = (100 \times 40.77) + (100 \times 0.1283) X_t$$

$$\hat{Y}_t^* = 4077 + 12.83 X_t$$

- 소득이 $1 증가하면, 식료품 지출은 12.83 센트 증가함
- 소득이 없는 사람은 식료품 지출에 4,077센트 사용함
- 모든 회귀계수 및 표준오차도 100배 증가
- 다른 모든 회귀분석 통계량(R^2, $t-$값 등)은 불변

✓ 모수에 대한 해석은 측정단위와 관련지어 이루어져야 함

- **회귀분석 결과, 그리고 수치들 사이의 상호관계**

Table 6.2 Computer Generated Least Squares Results

(1)	(2)	(3)	(4)	(5)
Variable	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob>|T|
INTERCEPT	40.7676	22.1387	1.841	0.0734
X	0.1283	0.0305	4.201	0.0002

- **회귀분석 결과, 그리고 수치들 사이의 상호관계(계속)**

$$b_1 = 40.7676 \qquad b_2 = 0.1283$$

$$se(b_1) = \sqrt{\hat{var}(b_1)} = \sqrt{490.12} = 22.1287$$

$$se(b_2) = \sqrt{\hat{var}(b_2)} = \sqrt{0.0009326} = 0.0305$$

$$t = \frac{b_1}{se(b_1)} = \frac{40.7676}{22.1287} = 1.84$$

$$t = \frac{b_2}{se(b_2)} = \frac{0.1283}{0.0305} = 4.20$$

- **회귀분석 결과, 그리고 수치들 사이의 상호관계(계속)**

Table 6.3 Analysis of Variance Table

Source	DF	Sum of Squares	Mean Square
Explained	1	25221.2229	25221.2229
Unexplained	38	54311.3314	1429.2455
Total	39	79532.5544	
		R-square	0.3171

- **회귀분석 결과, 그리고 수치들 사이의 상호관계(계속)**

$$SST = \sum(y_t - \bar{y})^2 = 79532$$

$$SSR = \sum(\hat{y}_t - \bar{y})^2 = 25221$$

$$SSE = \sum \hat{e}_t^{\,2} = 54311$$

$$SSE/(T\text{-}2) = \hat{\sigma}^2 = 1429.2455$$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 0.317$$

4.3.2 함수형태의 선택

- 단순**선형**회귀모형 $Y_t = \beta_1 + \beta_2 X_t + e_t$
- "선형"이라는 용어는 "변수들 사이의 선형"이 아니라 "모수들 사이의 선형"을 의미함
- 이 경우 "변수들 사이의 비선형" 모형은 적절한 변형을 통하여 "변수들 사이의 선형" 모형으로 전환될 수 있으며, OLS를 적용하여 추정할 수 있음

Linear vs. Nonlinear

Linear Statistical Models:

$$y_t = \beta_1 + \beta_2 x_t + e_t \qquad y_t = \beta_1 + \beta_2 \ln(x_t) + e_t$$

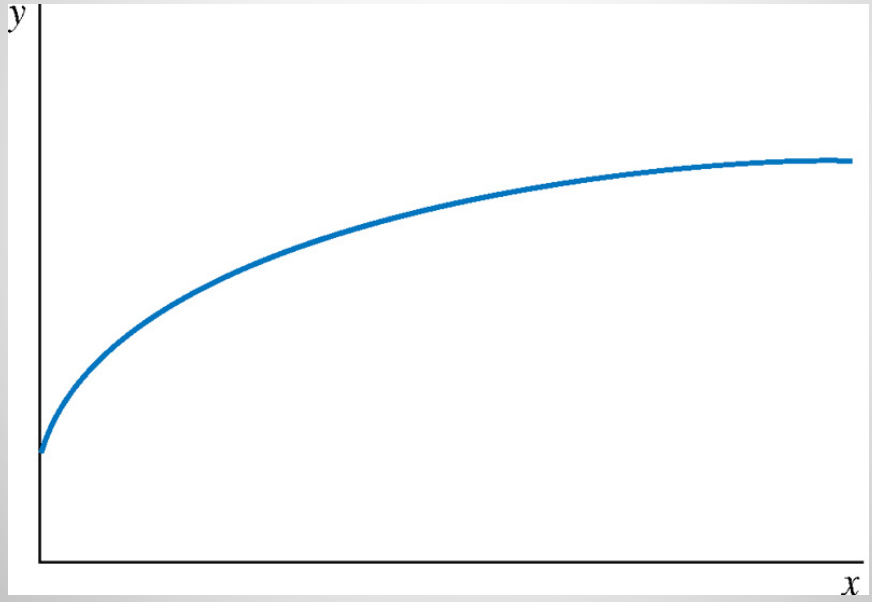
$$\ln(y_t) = \beta_1 + \beta_2 x_t + e_t \qquad y_t = \beta_1 + \beta_2 x_t^2 + e_t$$

Nonlinear Statistical Models:

$$y_t = \beta_1 + \beta_2 x_t^{\beta_3} + e_t \qquad y_t^{\beta_3} = \beta_1 + \beta_2 x_t + e_t$$

$$y_t = \beta_1 + \beta_2 x_t + \exp(\beta_3 x_t) + e_t$$

- **(예) 식료품 지출액**
- 아래와 같은 비선형관계일 가능성이 높음
- 선형모형으로 추정하면 적절치 못할 수 있음



- 변수간의 비선형관계

- 변수간의 비선형관계
- 자연대수: $\ln Y$, $\ln X$
- 역수: $1/Y$, $1/X$
- 제곱: Y^2, X^2
- 지수: e^Y, e^X

✓ 위와 같은 대수학적인 변형을 통하여
다양한 형태의 변수간 비선형관계를 나타낼 수 있음

- 비선형관계, 그렇지만 선형모형으로 변환 가능한 모형

Type	Statistical Model	Slope	Elasticity
1. Linear	$y_t = \beta_1 + \beta_2 x_t + e_t$	β_2	$\beta_2 \frac{x_t}{y_t}$
2. Reciprocal	$y_t = \beta_1 + \beta_2 \frac{1}{x_t} + e_t$	$-\beta_2 \frac{1}{x_t^2}$	$-\beta_2 \frac{1}{x_t y_t}$
3. Log-Log	$\ln(y_t) = \beta_1 + \beta_2 \ln(x_t) + e_t$	$\beta_2 \frac{y_t}{x_t}$	β_2
4. Log-Linear (Exponential)	$\ln(y_t) = \beta_1 + \beta_2 x_t + e_t$	$\beta_2 y_t$	$\beta_2 x_t$
5. Linear-Log (Semi-log)	$y_t = \beta_1 + \beta_2 \ln(x_t) + e_t$	$\beta_2 \frac{1}{x_t}$	$\beta_2 \frac{1}{y_t}$
6. Log-inverse	$\ln(y_t) = \beta_1 - \beta_2 \frac{1}{x_t} + e_t$	$\beta_2 \frac{y_t}{x_t^2}$	$\beta_2 \frac{1}{x_t}$

- **함수형태의 선택: 실증적인 논의**

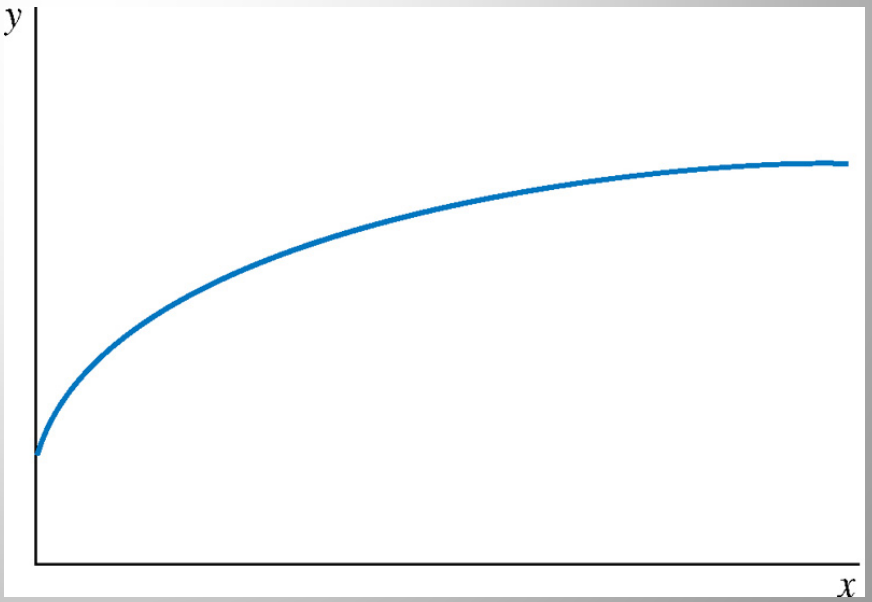
- 최선의 함수형태를 어떻게 고를 수 있을까?
  - 모든 상황에 적용될 수 있는 원칙은 없음

- 실증분석을 할 때 유용한 방법들
  - 경제이론이 의미하는 함수형태
  - 단순회귀모형의 기본가정이 준수되는 함수형태
  - 자료의 산포도를 그려보고 그것을 잘 반영할 수 있는 함수형태

- **식료품 지출액 사례의 모형화 방법**

$$FOOD\_EXP = \beta_1 + \beta_2 \frac{1}{INCOME} + e$$

$$FOOD\_EXP = \beta_1 + \beta_2 \ln(INCOME) + e$$



▪ 함수형태를 선택하는 기준

- 만족스러운 함수형태를 선택할 경우,
  회귀모형의 기본 가정이 준수되는데 도움이 됨
- 함수형태를 선택하거나 변수를 변형시키는 주요한 목적은
  오차항이 다음과 같은 특성을 갖는 모형을 설정하기 위한 것.
  이 경우에 OLS 추정량은 타당한 통계적 특성을 가짐

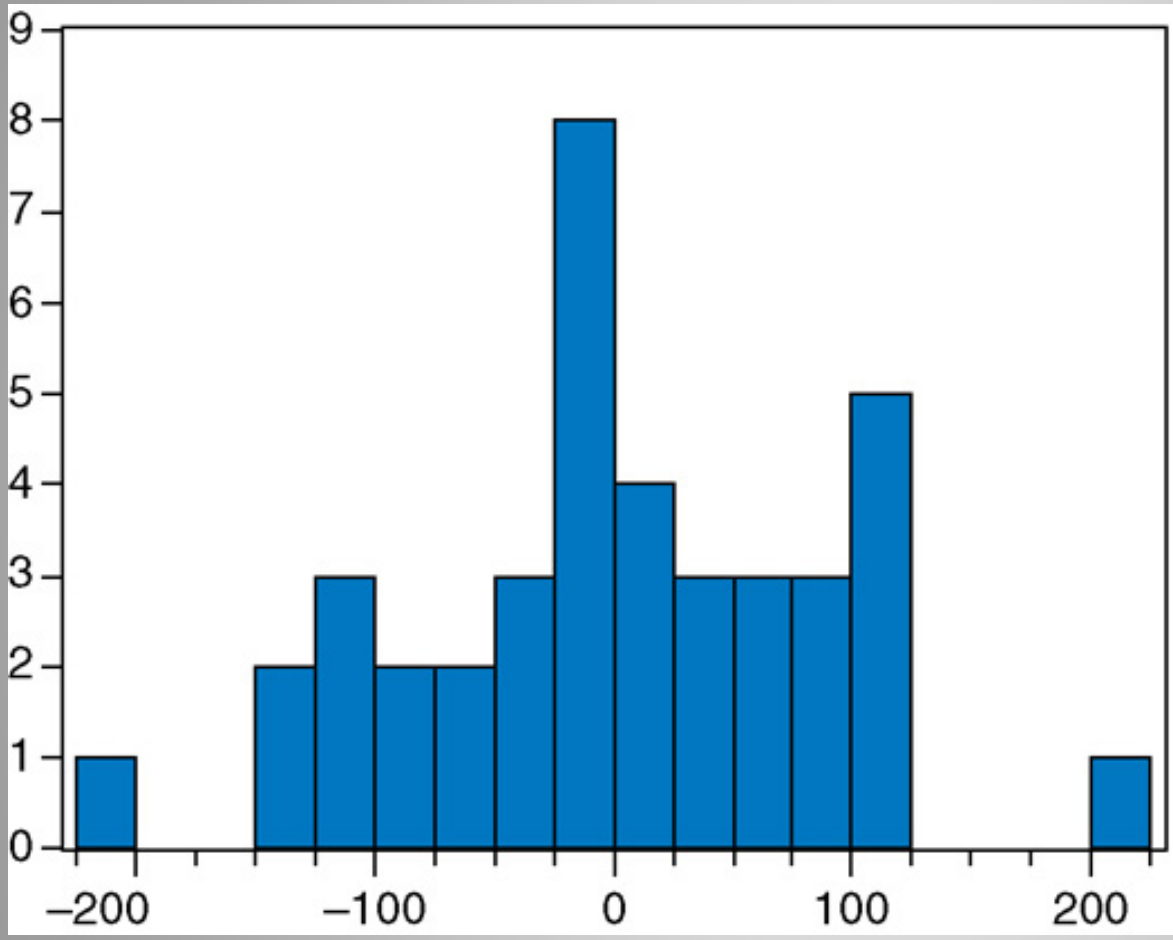
1. $E(e_t)=0$
2. $\mathrm{var}(e_t)=\sigma^2$
3. $\mathrm{cov}(e_i,e_j)=0$
4. $e_t \sim N(0, \sigma^2)$

※ R^2 가 높다고 꼭 좋은 모형은 아님

4.3.4 잔차는 정규분포를 하는가?

- OLS 추정량을 이용한 가설검정과 구간추정은
  잔차가 정규분포를 한다는 가정에 의존함.
  따라서 함수형태를 선택할 때,
  잔차가 정규분포를 하는 것이 바람직함
- 잔차의 정규성(normality)을 검정하는 통계량들이
  대부분의 컴퓨터 SW에서 계산되어 제공됨
  (예) J-B 통계량

4.3.4 잔차는 정규분포를 하는가?



Series: Residuals	
Sample 1 40	
Observations 40	
Mean	6.93e-15
Median	–6.324473
Maximum	212.0440
Minimum	–223.0255
Std. Dev.	88.36190
Skewness	–0.097319
Kurtosis	2.989034
Jarque–Bera	0.063340
Probability	0.968826

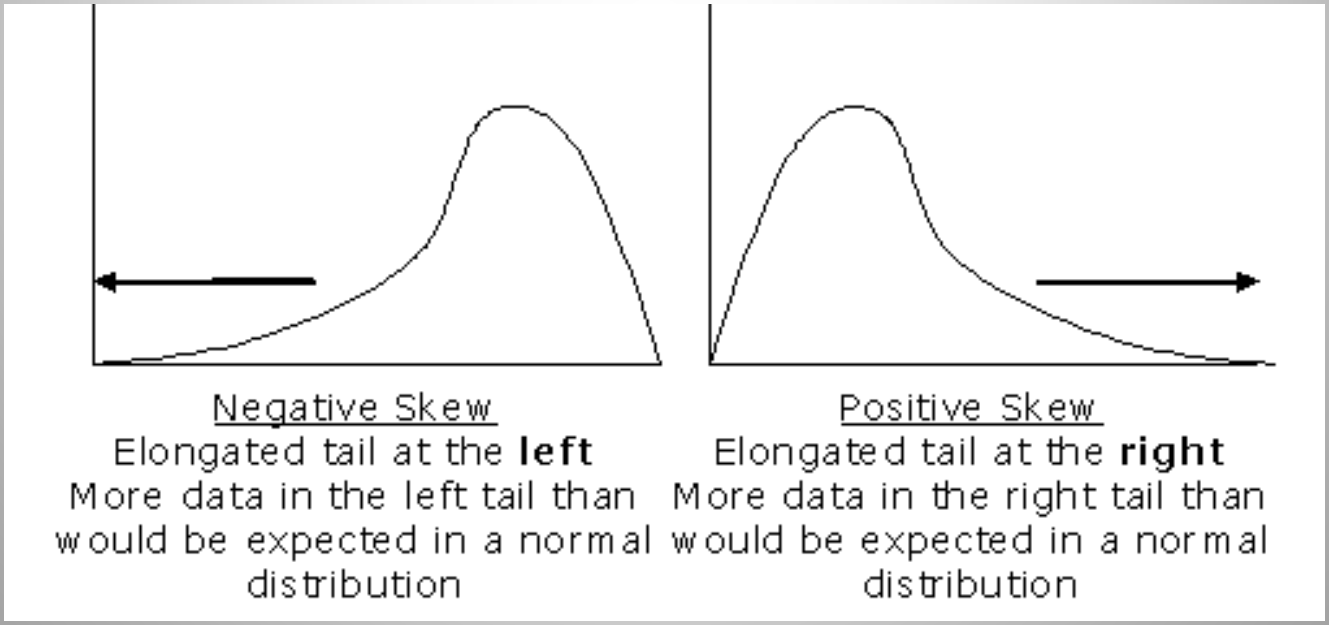
- **쟈크-베라 검정 (Jarque-Bera test, J-B test)**

- 귀무가설: "잔차는 정규분포를 한다"

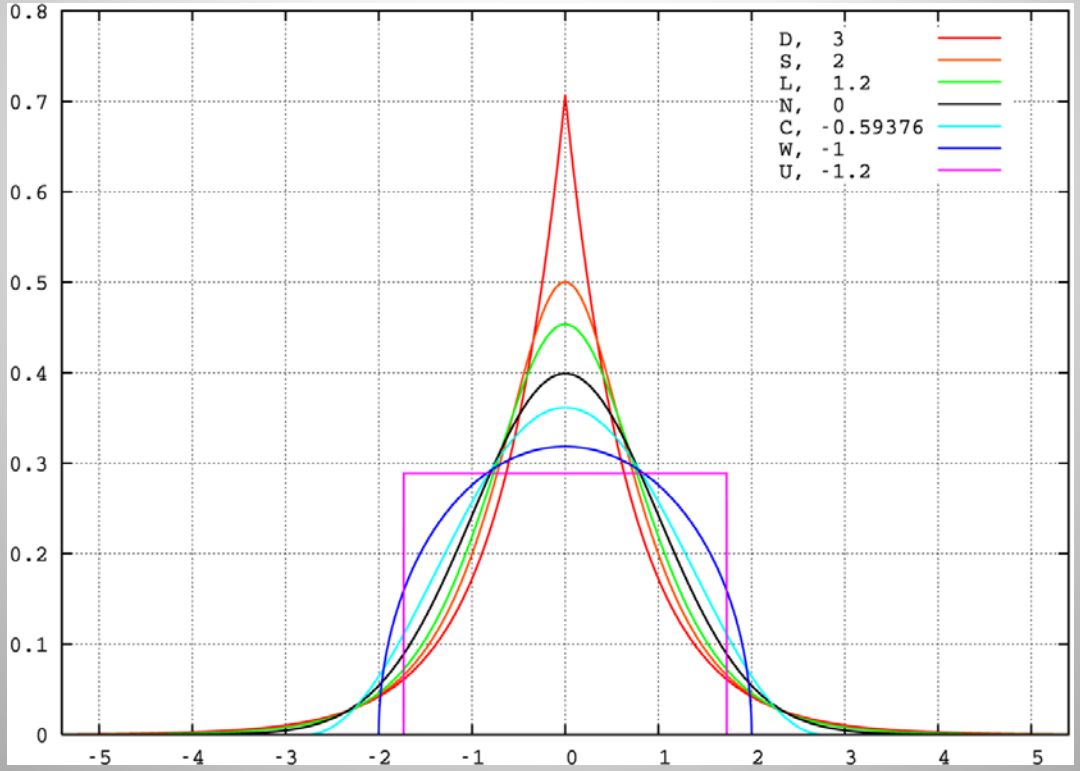
$$JB = \frac{T}{6}\left(s^2 + \frac{(k-3)^2}{4} \right) \sim \chi^2(2)$$

- s 는 왜도(비대칭도, skewness)
- k 는 첨도(첨예도, kurtosis)

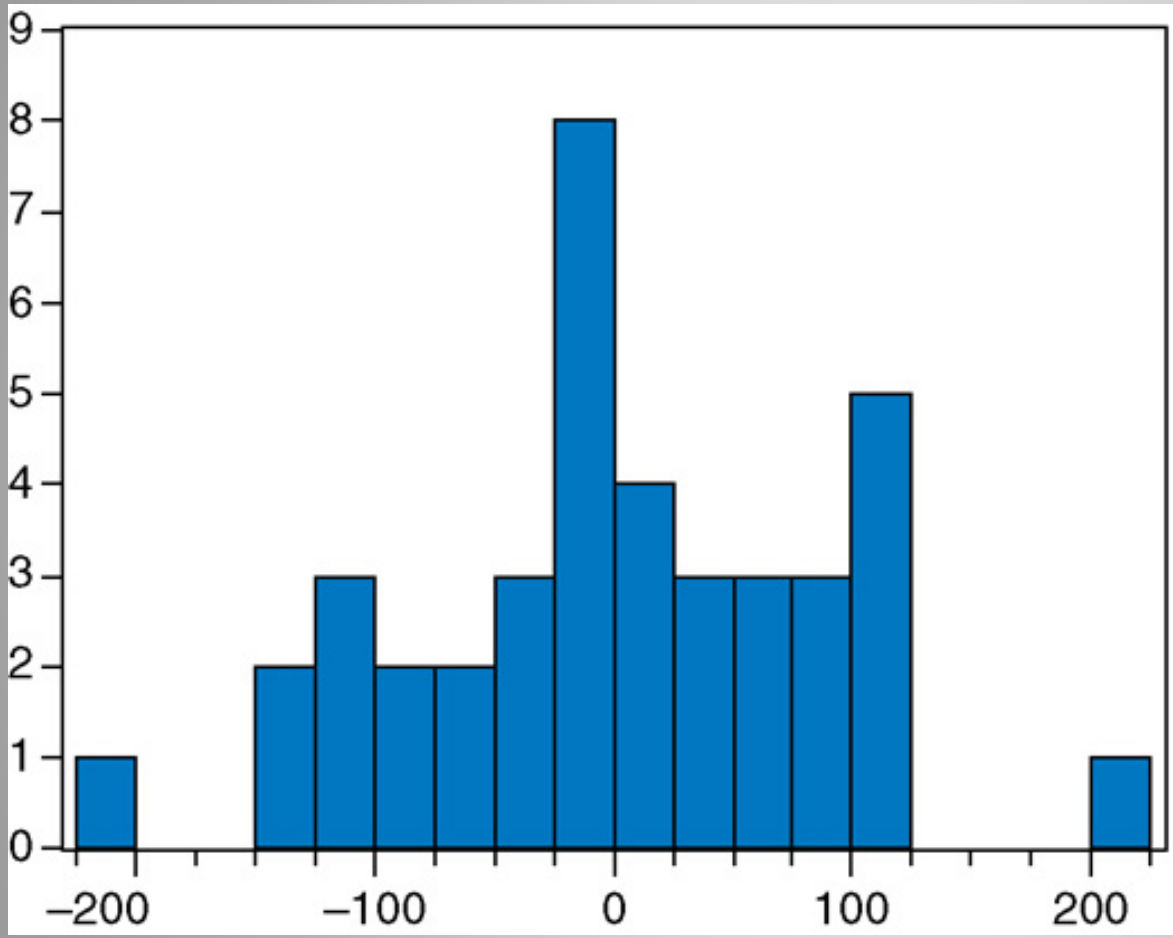
- **왜도 (skewness)**
- 분포의 비대칭도를 나타냄
- 정규분포에서는 $s = 0$
- $s < 0$, 왼쪽꼬리분포; $s > 0$, 오른쪽꼬리분포



- **첨도 (kurtosis)**
- 분포의 첨예도, 자료가 얼마나 뾰족한가를 판단하는 척도
- 정규분포에서는 $k = 3$ (execss kurtosis = 0)
- 첨도가 클수록 분포는 뾰족한 모습



- **잔차는 정규분포를 하는가?**



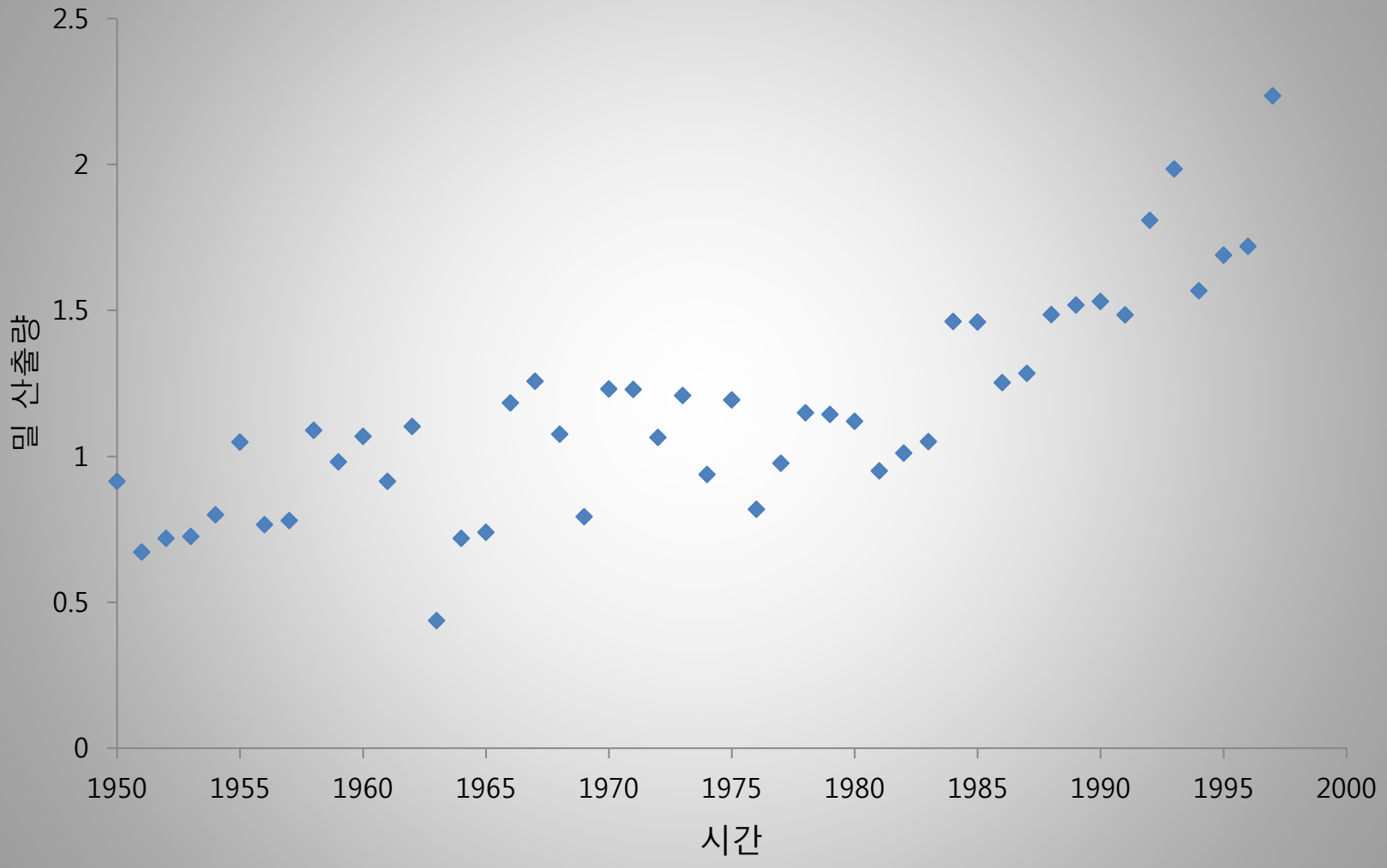
Series: Residuals	
Sample 1 40	
Observations 40	
Mean	6.93e-15
Median	–6.324473
Maximum	212.0440
Minimum	–223.0255
Std. Dev.	88.36190
Skewness	–0.097319
Kurtosis	2.989034
Jarque–Bera	0.063340
Probability	0.968826

- **식료품비 지출액 잔차의 정규성 검정**

$$JB = \frac{40}{6}\left(-0.097^2 + \frac{(2.99-3)^2}{4} \right) = 0.063$$

- $\chi^2(2)$ 의 5% 임계값은 5.99 (1% 임계값은 9.21)
- 0.063<5.99 이므로,
  잔차가 정규분포를 한다는 귀무가설을 기각할 수 없음
- p-값=0.9688>0.05(유의수준), 귀무가설을 기각할 수 없음

4.3.5 밀 산출량 산포도 (p.143 사례)



▪ 선형모형으로 추정

$$Y_t = \beta_1 + \beta_2 X_t + e_t$$

$$\hat{Y}_t = 0.638 + 0.021 X_t \qquad R^2 = 0.649$$

$$(0.064)\ \ (0.002) \qquad (s.e.)$$

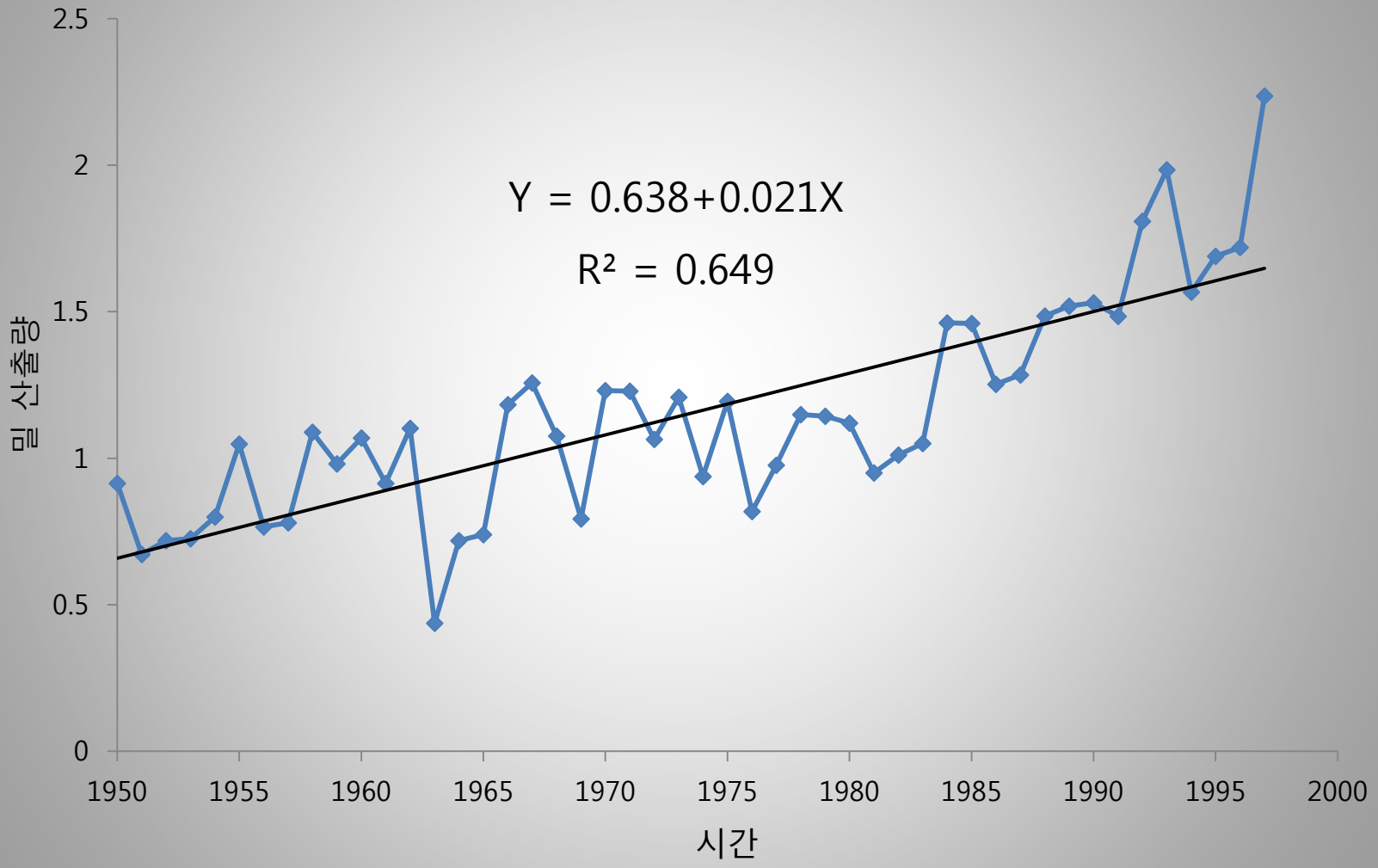
회귀분석 통계량	
다중 상관계수	0.80585
결정계수	0.649394
조정된 결정계수	0.641772
표준 오차	0.218692
관측수	48

분산 분석

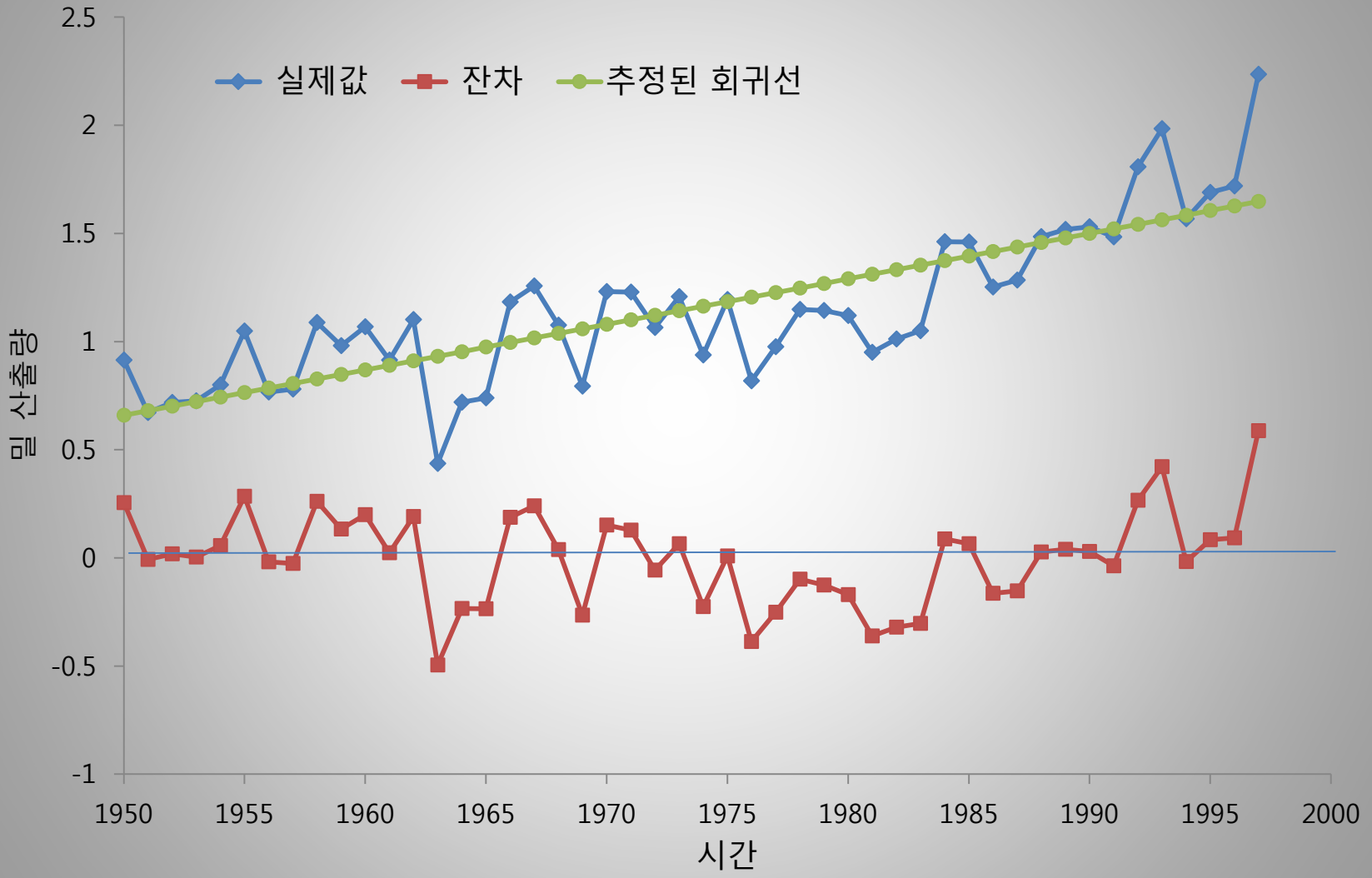
	자유도	제곱합	제곱 평균	F 비	유의한 F
회귀	1	4.07486	4.07486	85.20125	4.88E-12
잔차	46	2.200009	0.047826		
계	47	6.274869			

	계수	표준 오차	t 통계량	P-값
Y 절편	0.637778	0.064131	9.944999	4.85E-13
X 1	0.021032	0.002279	9.230452	4.88E-12

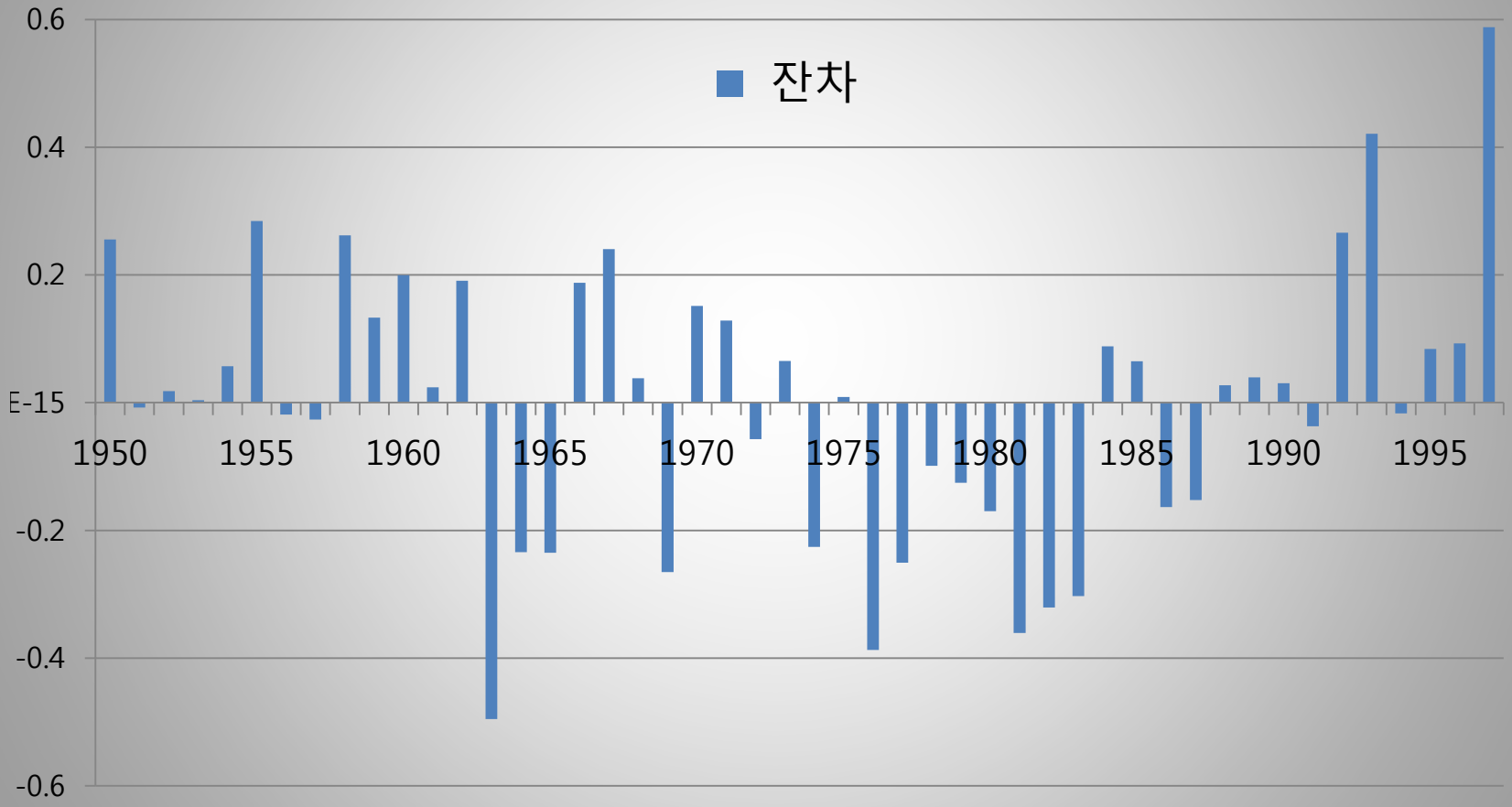
▪ 밀 산출량 표본회귀선의 추정



▪ 잔차 계산


실제값
잔차
추정된 회귀선
밀 산출량
시간
2.5
2
1.5
1
0.5
0
-0.5
-1
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000

- 잔차가 (+), (-), (+) 형태로 집중되어 있음
  밀 산출량의 증가 속도가 빠르다는 의미, 비선형 모형 필요함



▪ 비선형모형으로 추정

$$Y_t = \beta_1 + \beta_2 X_t^3 + e_t$$

$$\hat{Y}_t = 0.874 + 9.682 Z_t^3 \qquad R^2 = 0.751$$

$$(0.036)\ \ (0.824) \qquad (s.e.)$$

$$Z_t^3 = X_t^3 / 1{,}000{,}000$$

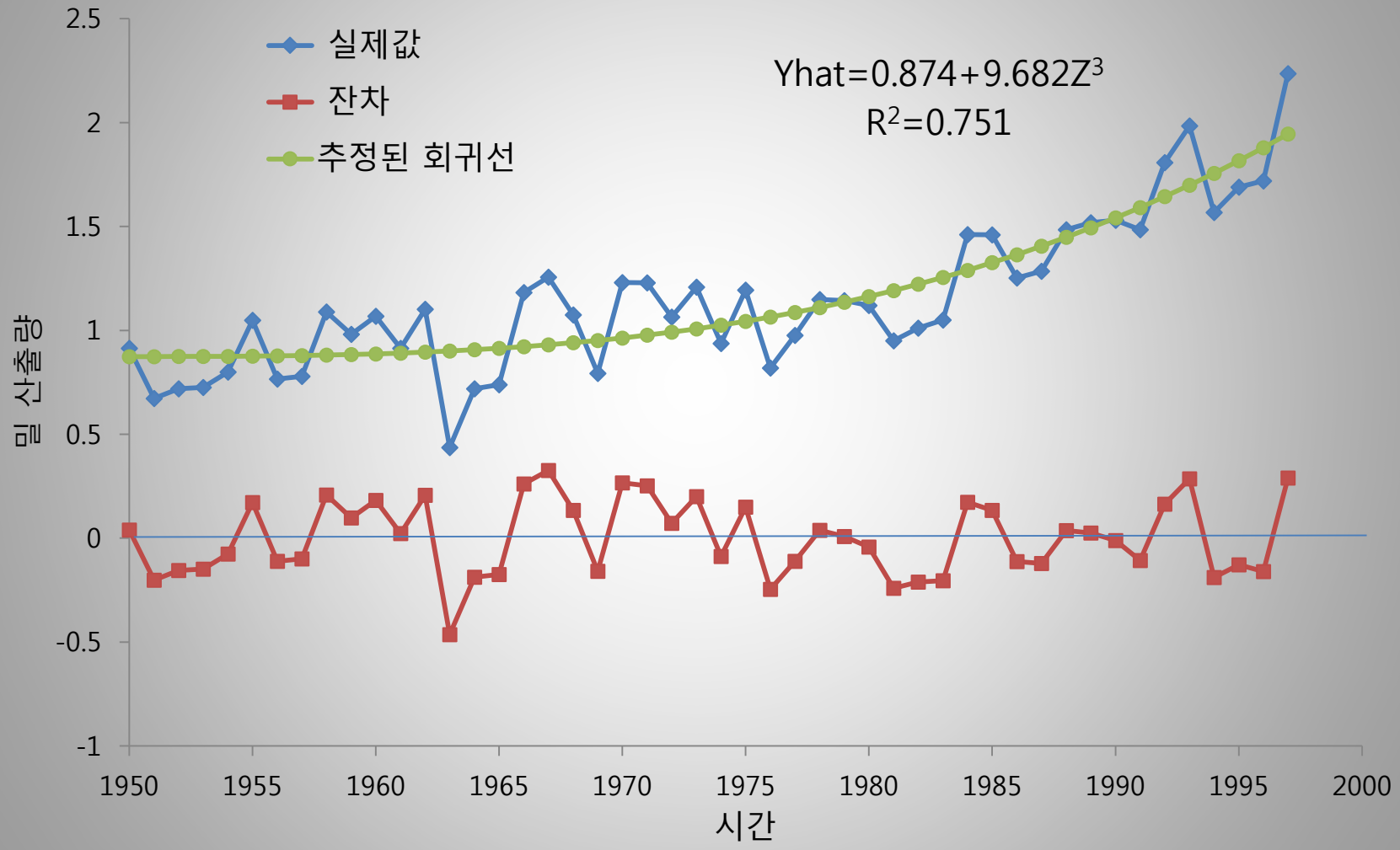
회귀분석 통계량	
다중 상관계수	0.866496
결정계수	0.750815
조정된 결정계수	0.745398
표준 오차	0.184368
관측수	48

분산 분석

	자유도	제곱합	제곱 평균	F 비	유의한 F
회귀	1	4.711265	4.711265	138.6017	1.77E-15
잔차	46	1.563604	0.033991		
계	47	6.274869			

	계수	표준 오차	t 통계량	P-값
Y 절편	0.874117	0.035631	24.5327	4.6E-28
X 1	9.681516	0.822355	11.77292	1.77E-15

- **잔차의 집중화 개선, R^2 증가**



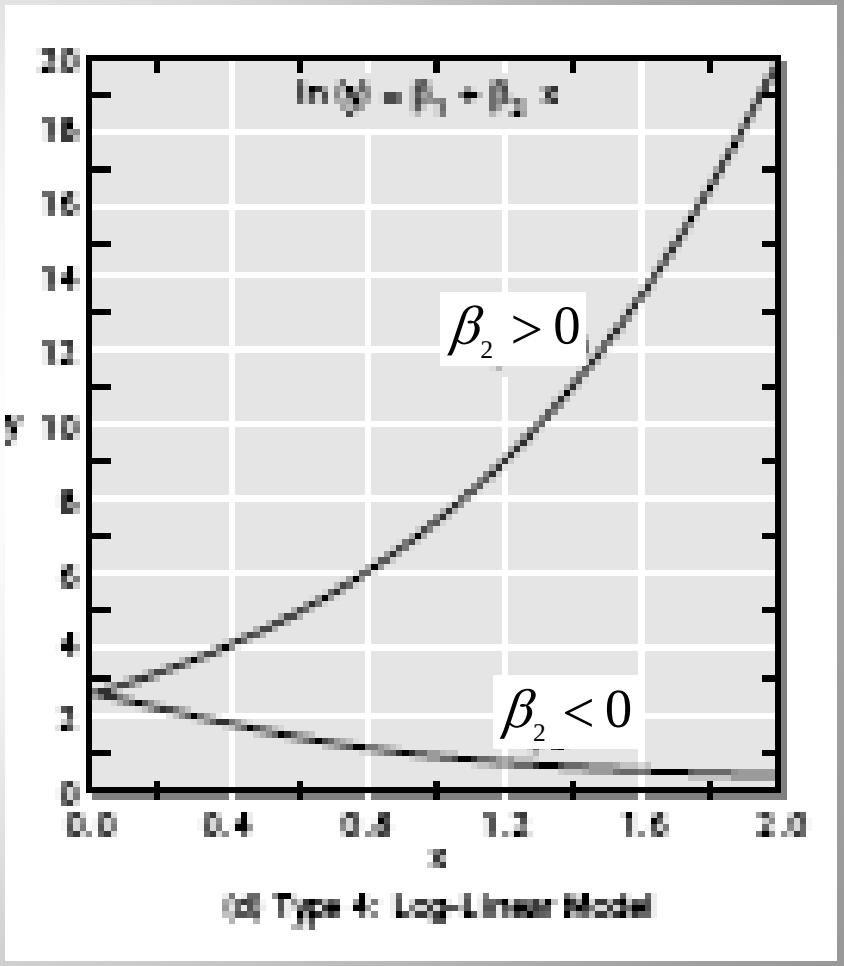
4.4 대수-선형모형(log-linear model)

$$\ln Y_t = \beta_1 + \beta_2 X_t + e_t$$

- Slope: $\beta_2 Y_t$
- *if* $\beta_2 > 0$, 체증하는 율로 증가
- Elasticity: $\beta_2 X_t$
- 해석:

  X 가 1단위 증가할 때,

  Y 는 대략 $100\beta_2\%$ 변화함



4.4.1 성장모형

- 앞의 밀 생산 사례에서 기술이 진보함에 따라(즉, 시간이 흐름에 따라) 밀 생산은 체증하는 율로 증대되었음
- 밀 생산량이 연간 일정한 율(g)로 증가한다면,

$$YIELD_t = (1+g)YIELD_{t-1}$$

- 이런 관계를 반복적으로 대체시키면,

$$\ln(YIELD_t) = \ln(YIELD_0) + \ln(1+g)t$$

$$= \beta_1 + \beta_2 t$$

- 여기서 $YIELD_0$ 는 0연도의 생산량(몰라도 좋음, 추정 가능)

▪ 밀 생산 사례: 대수-선형 모형

- 밀 생산 통계자료를 대수-선형 모형으로 추정해보면,

$$\widehat{\ln\left(YIELD_t\right)} = -0.3434 + 0.0178t$$
$$(\text{se}) \qquad (0.0584) \quad (0.0021)$$

- 추정치 $b_2 = \widehat{\ln\left(1+g\right)} = 0.0178$
- x 가 작은 경우 $\ln(1+x) \cong x$ 라는 사실을 이용하면,
- 밀 생산의 증가율은 대략 $\hat{g} = 0.0178$

  또는 연간 약 1.78%로 추정됨

  $100\beta_2\%$

4.4.2 임금방정식

- 교육을 추가적으로 1년 더 받으면, 임금이 일정 율(r)로 증가한다면

$$WAGE_1 = (1+r)WAGE_0$$

$$WAGE_2 = (1+r)WAGE_1 = (1+r)^2 WAGE_0$$

- 교육 받은 기간(년)을 $EDUC$ 라고 나타내면, 다음이 성립함

$$\ln(WAGE) = \ln(WAGE_0) + \ln(1+r)EDUC$$

$$= \beta_1 + \beta_2 EDUC$$

- 교육을 추가적으로 1년 더 받으면, 임금은 대략 $100\beta_2\%$ 증가

- **미국 통계를 이용한 임금방정식 추정결과**

$$\widehat{\ln(WAGE)} = 0.7884 + 0.1038 \times EDUC$$
$$(\text{se}) \qquad (0.0849) \quad (0.0063)$$

- 교육을 추가적으로 1년 더 받으면, 임금은 대략 10.4% 증가함

4.4.3 대수-선형 모형에서의 예측

- 소표본 경우: $\hat{y}_n = \exp\left(\widehat{\ln(y)}\right) = \exp(b_1 + b_2 x)$
- 대표본 경우: $\hat{y}_c = \widehat{E(y)} = \exp\left(b_1 + b_2 x + \hat{\sigma}^2/2\right) = \hat{y}_n e^{\hat{\sigma}^2/2}$

(교과서 부록 4C에 증명 참조)

- 교육받은 기간이 12년인 근로자의 임금은 얼마인가?

$$\widehat{\ln(WAGE)} = .7884 + .1038 \times EDUC = .7884 + .1038 \times 12 = 2.0335$$

$$\hat{y}_{12} = \exp\left(\widehat{\ln(y)}\right) = \exp(2.0335) = 7.6408 \quad \text{(시간당 7.64 달러)}$$

- 수정된 예측값은 $\hat{y}_c = \widehat{E(y)} = \hat{y}_n e^{\hat{\sigma}^2/2} = 7.6408 \times 1.1276 = 8.6161$
- (시간당 8.61 달러)

4.4.4 일반화된 R^2의 측정

$$R_g^2 = \left[\text{corr}\left(y, \hat{y}\right)\right]^2 = r_{y,\hat{y}}^2$$

- 임금방정식의 사례

$$\widehat{\ln\left(WAGE\right)} = 0.7884 \; + \; 0.1038 \times EDUC \qquad R^2 = 0.2146$$

$$(\text{se}) \qquad (0.0849) \quad (0.0063)$$

$$R_g^2 = \left[\text{corr}\left(y, \hat{y}_c\right)\right]^2 = 0.4739^2 = 0.2246$$

❖ 대수-대수모형(log-log model)

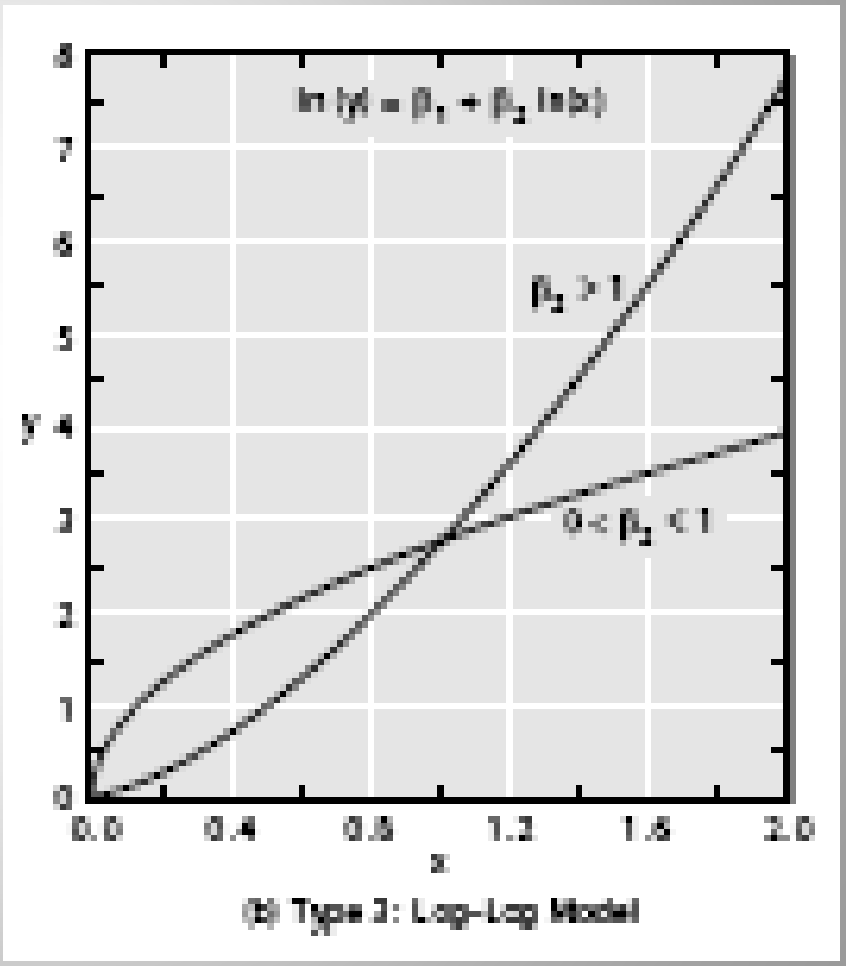
$$\ln Y_t = \beta_1 + \beta_2 \ln X_t + e_t$$

- Slope

$$\beta_2 \frac{Y_t}{X_t}$$

- Elasticity

$$\beta_2$$



❖ 대수-대수모형(log-log model)-계속

$$\ln Y_t = \beta_1 + \beta_2 \ln X_t + e_t$$

- Slope

$$\beta_2 \frac{Y_t}{X_t}$$

- Elasticity

$$\beta_2$$

