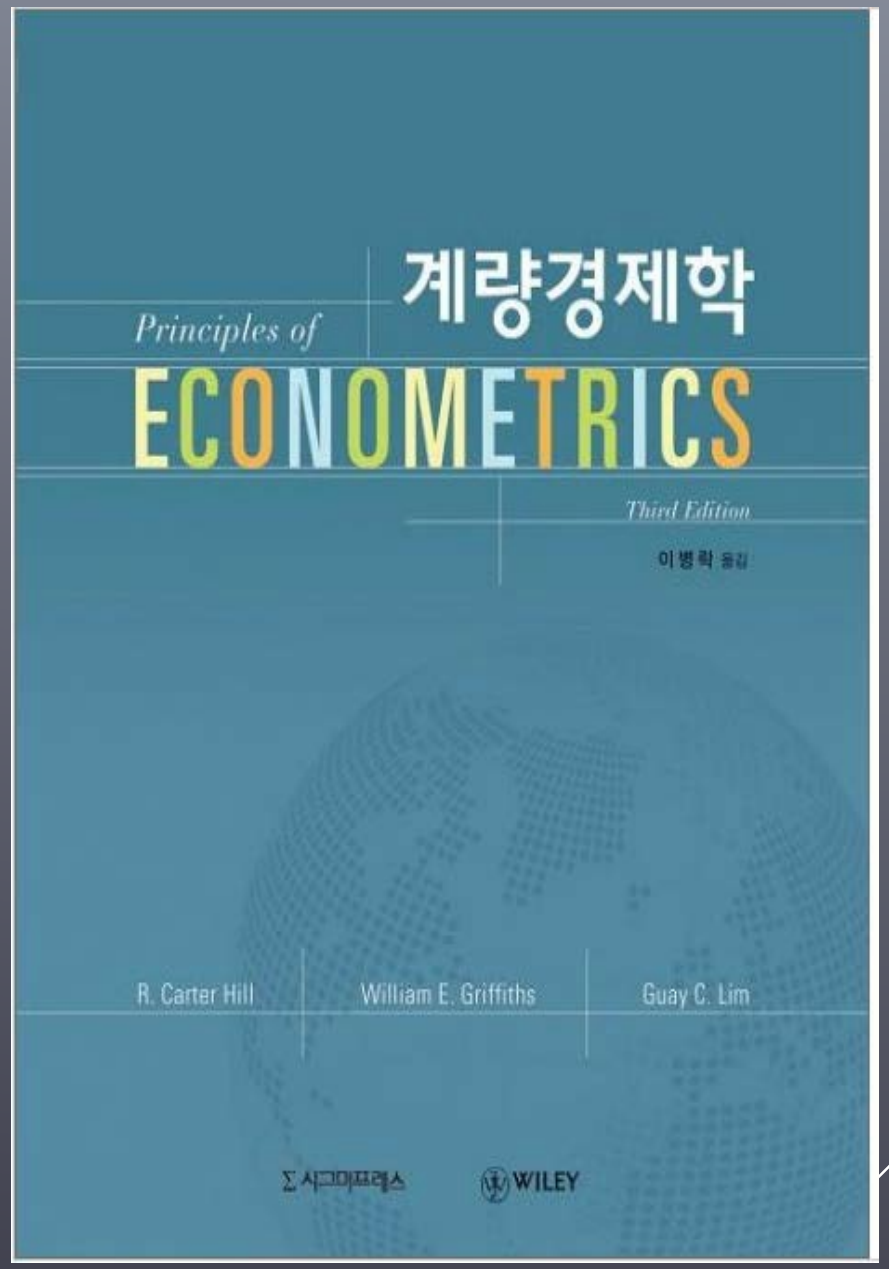


제2장
단순 선형회귀 모형

0. 상관분석과 회귀분석

- 상관분석 – 단순한 양적 관련성 분석
  회귀분석 – **인과관계** 분석

- **Purpose of Regression Analysis**

1. **Estimate a relationship** among economic variables, such as y = f(x), **quantitatively**.
2. **Testing hypothesis**
3. **Forecast** or **predict** the value of one variable, y, based on the value of another variable, x.

1. 경제모형

- 경제적 인과관계를 나타내는 함수로 나타낼 수 있음
- x : 원인, y :결과라면, 경제모형은 다음과 같이 표현됨

implicit functional form: $y = f(x)$

explicit functional form: $E(y \mid x) = \beta_1 + \beta_2 x$

$$E(y \mid x) = \beta_1 + \beta_2 \ln x$$

$$\ln E(y \mid x) = \beta_1 + \beta_2 \ln x$$

$$E(y \mid x) = \beta_1 + \beta_2 e_x$$

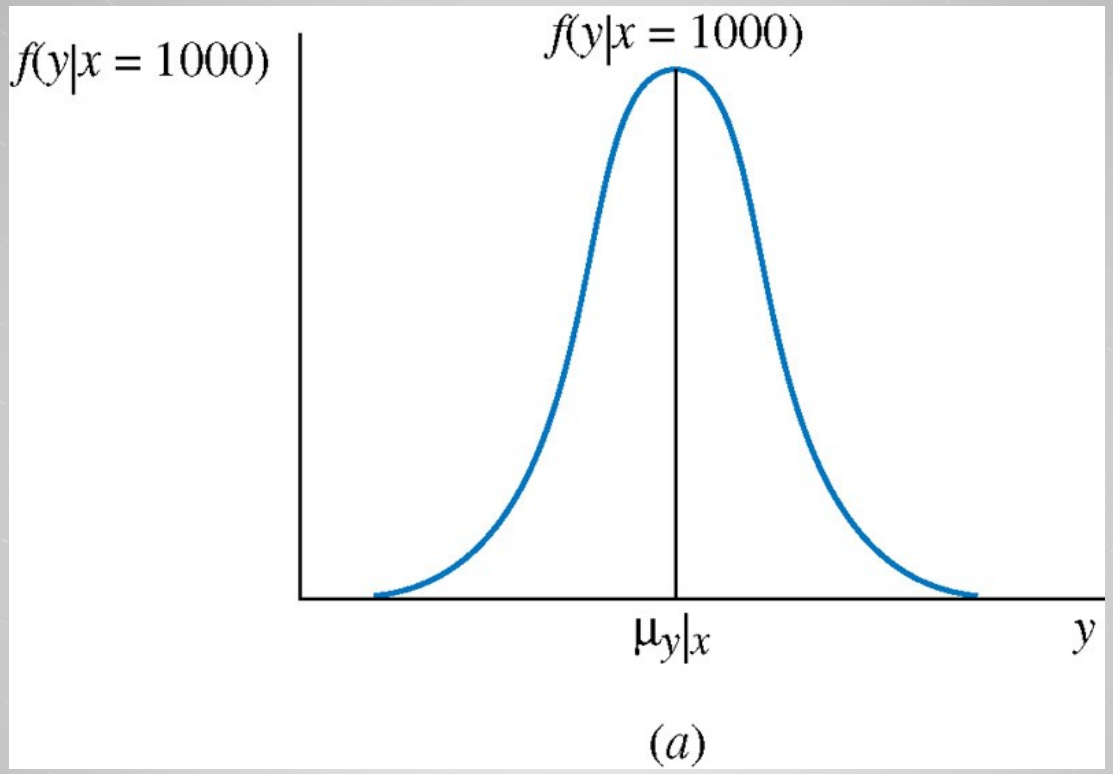
Example: **Weekly Food Expenditures**

y = dollars spent each week on food items
x = consumer's weekly income

- The relationship between x and the expected value of y, given x, might be **linear:**

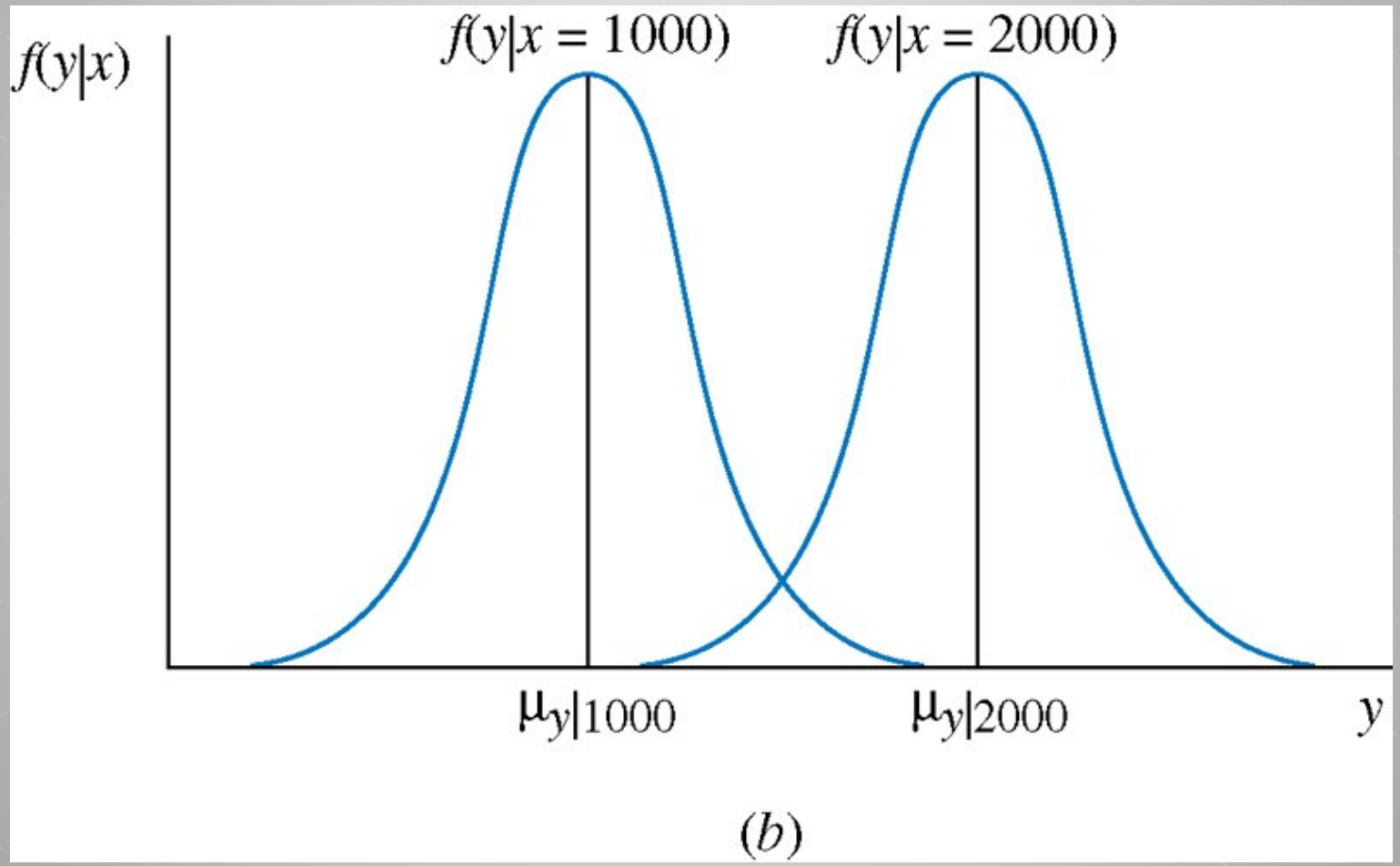
$$E(y \mid x) = \beta_1 + \beta_2 x$$

- 가계소득이 주당 $1,000인 가계만 조사했다고 해보자



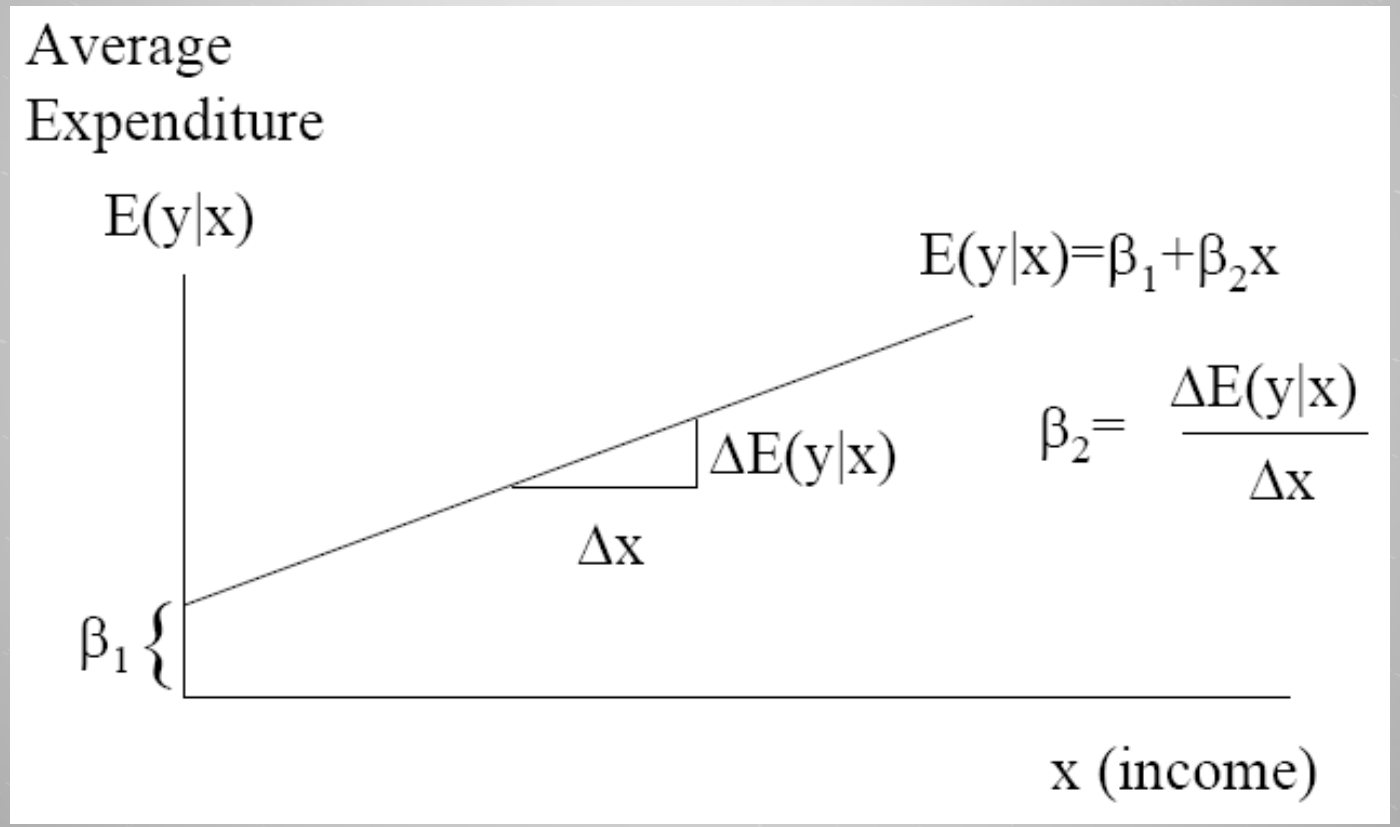
- 다수의 가계가 평균수준의 식료품비를 지출하지만, 평균 이상 혹은 이하를 지출하는 가계도 많이 있음
  ⇒ 조건부 식료품비 지출액은 정규분포를 함

- Probability distribution of food expenditures if given income x=$1,000 and x=$2,000.



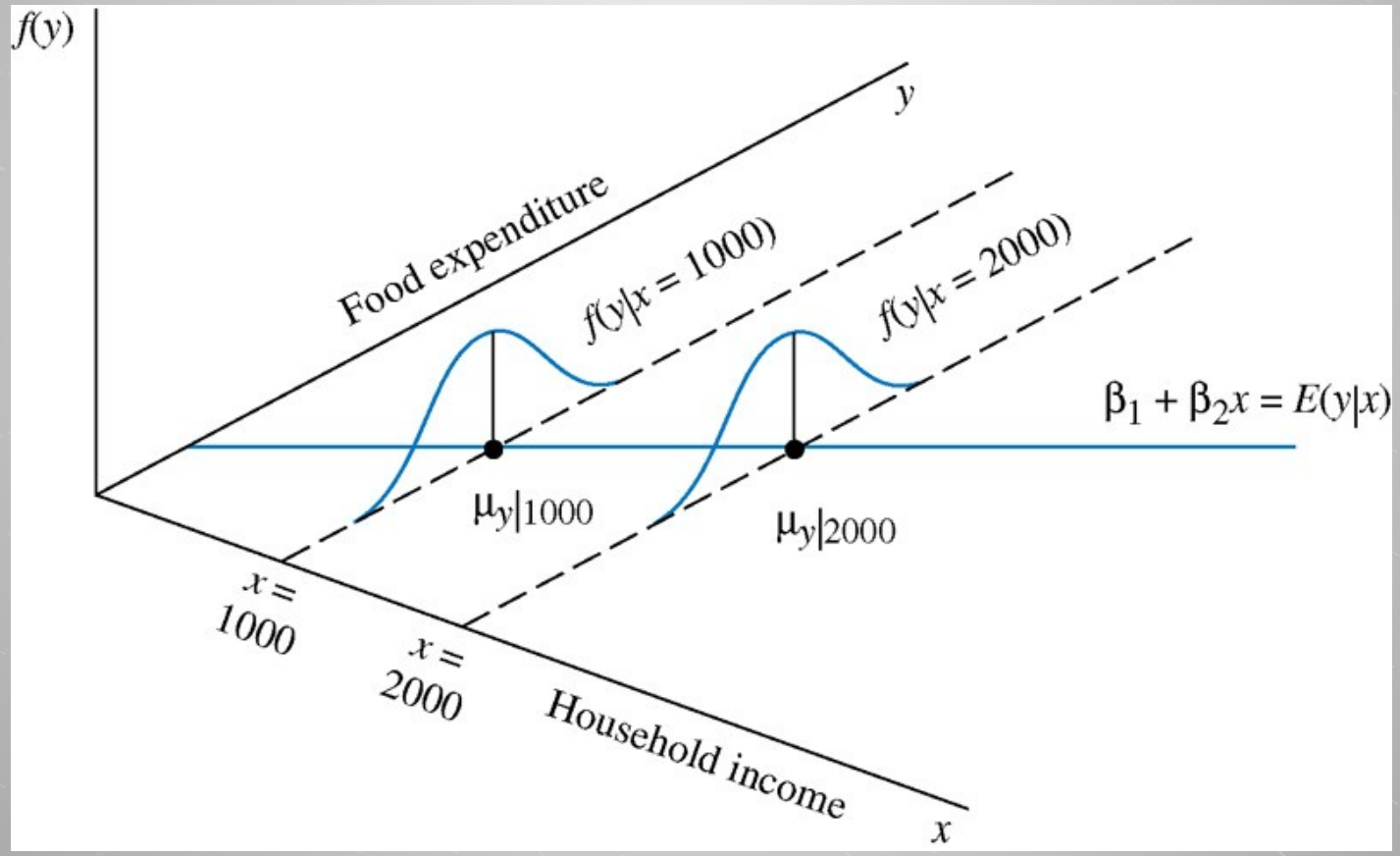
- **economic model**

  : a linear relationship between
  average expenditure on food and income



2. 계량경제모형 (econometric model)

- 실제 지출액은 평균값을 중심으로 분포되어 있음
- 실제 지출액의 평균값을 연결하면 "모회귀선"이 됨



- 경제모형: 경제변수 사이의 평균적 관계를 나타냄

$$E(y \mid x) = \beta_1 + \beta_2 x$$

- 계량경제모형: 경제변수 사이의 무작위적 관계도 나타냄

$$y = E(y \mid x) + e \qquad y = \beta_1 + \beta_2 x + e$$

평균적 혹은
체계적 관계

무작위적 요소
= 무작위 오차항
(random error term)

계량경제모형의 기본가정(Ⅰ)

- 각 소득수준(x)에 대한 평균 가계 지출액은 다음과 같이 나타낼 수 있음 $E(y|x)=\beta_1+\beta_2 x$
- 각 가계의 지출액은 각 소득수준에서 $E(y|x)=\beta_1+\beta_2 x$ 주변에 분산되어 흩어져 있음

  이 흩어진 (혹은 퍼진) 정도는 동일하다고 가정함
- 각각의 표본은 무작위적이라고 생각함
- x 가 최소한 두 개의 다른 값을 가져야 함
- y 의 분포, $f(y|x)$ 는 정규분포를 함 (선택적 가정)

Assumptions of the Simple Linear Regression Model - I

1. The average value of y, given x, is given by the **linear** regression:

$$\mathbf{E(y) = \beta_1 + \beta_2 x}$$

2. For each value of x, the values of y are distributed around their mean with **variance**:

$$\mathbf{var(y) = \sigma^2}$$

3. The values of y are uncorrelated, having **zero covariance** and thus no linear relationship:

$$\mathbf{cov(y_i, y_j) = 0}$$

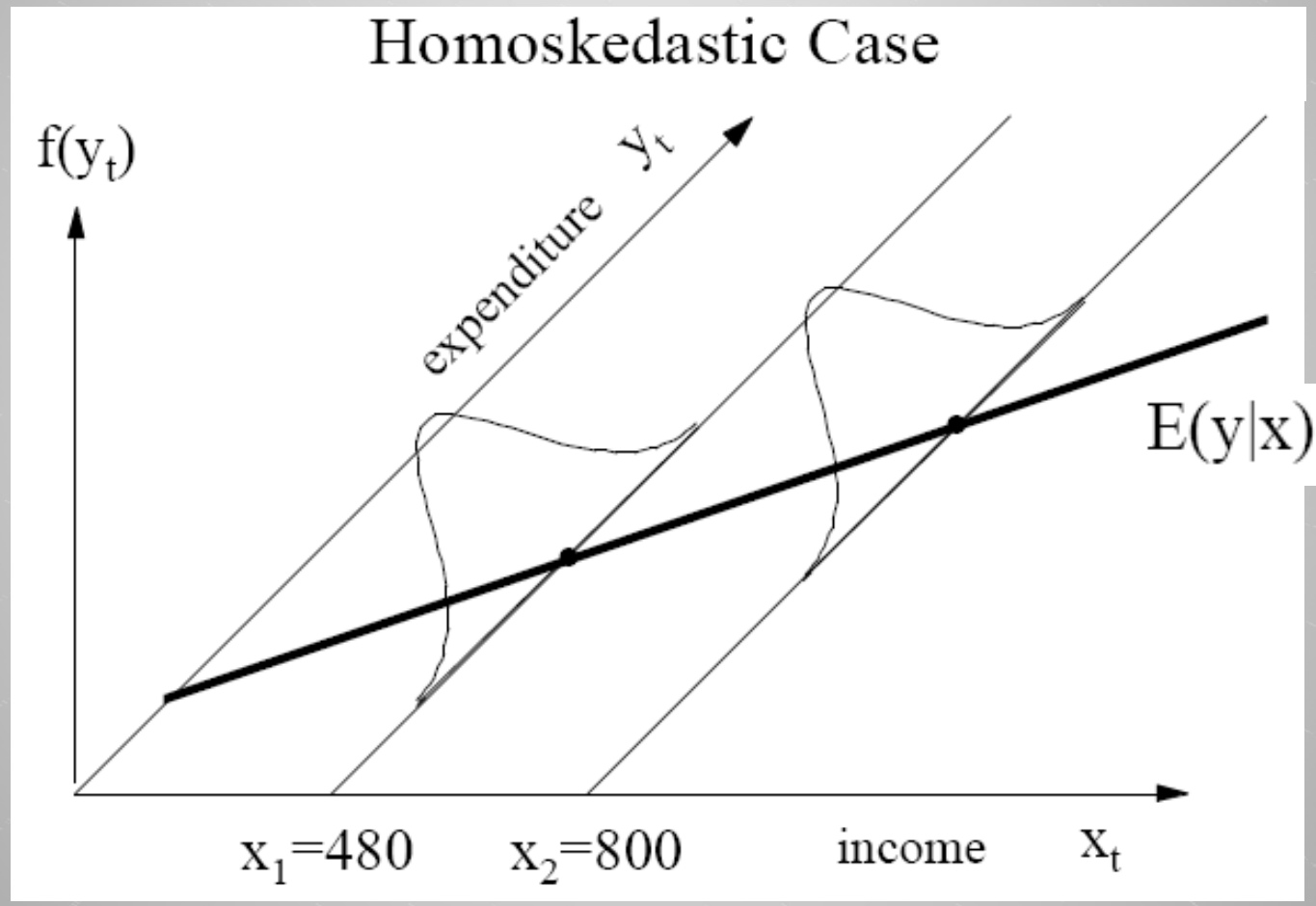
4. The variable x must take **at least two different values**, so that $\mathbf{x \neq c}$, where c is a constant.

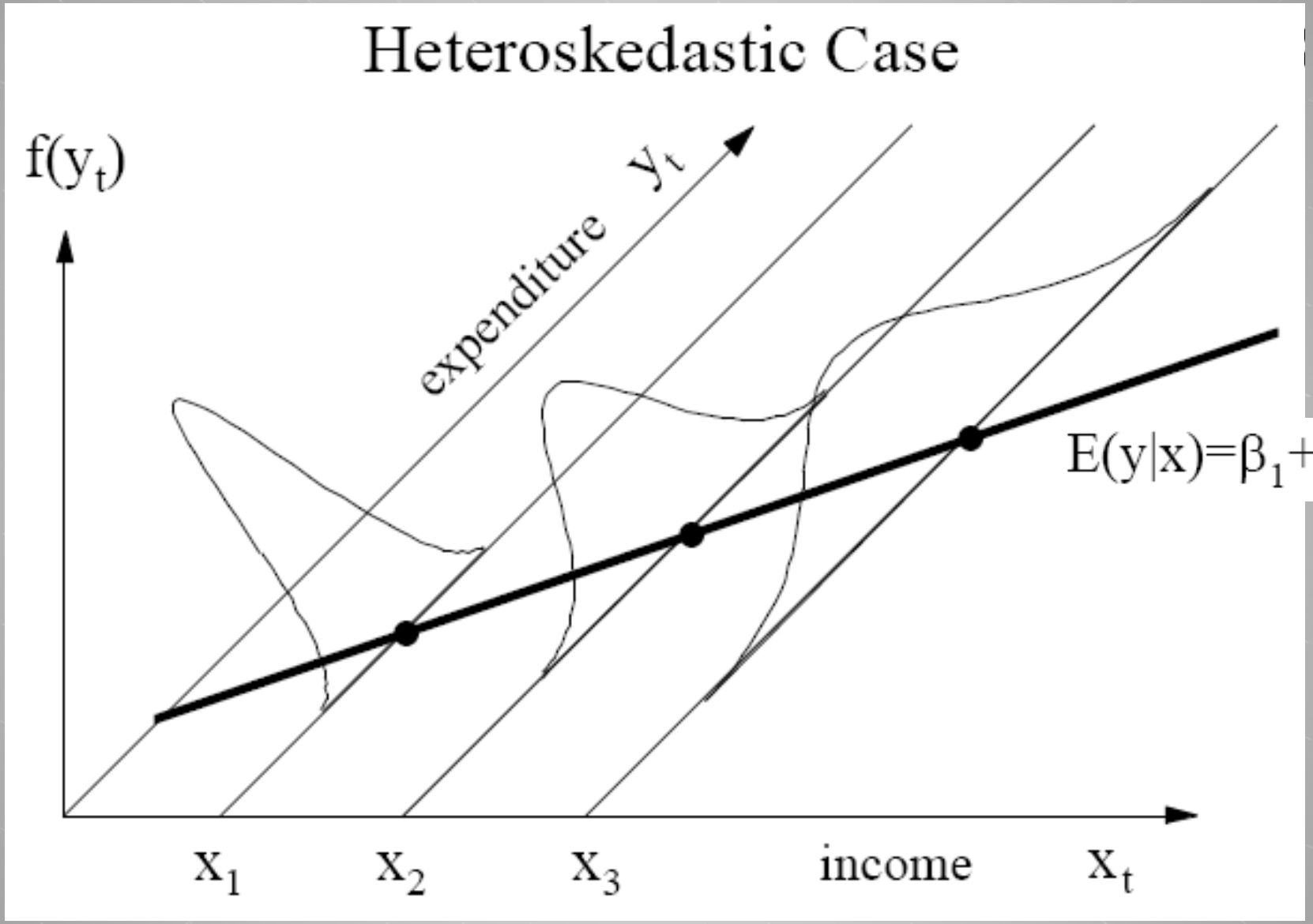
One more assumption that is often used in practice but is not required for least squares:

5. (optional) The values of y are **normally distributed** about their mean for each value of x:

$$y \sim N\left[(\beta_1+\beta_2 x), \sigma^2\right]$$

- The probability density function for y_t at two levels of household income, x_t




Heteroskedastic Case
f(y_t)
y_t
expenditure
E(y|x)=β_1+β_2x
x_1
x_2
x_3
income
x_t

▪ 오차항 (error term) 혹은 교란항 (disturbance term)

- y is a random variable composed of **two parts**:

① **Systematic component**: $E(y) = E(y \mid x) = \beta_1 + \beta_2 x$

This is the **mean of y.**

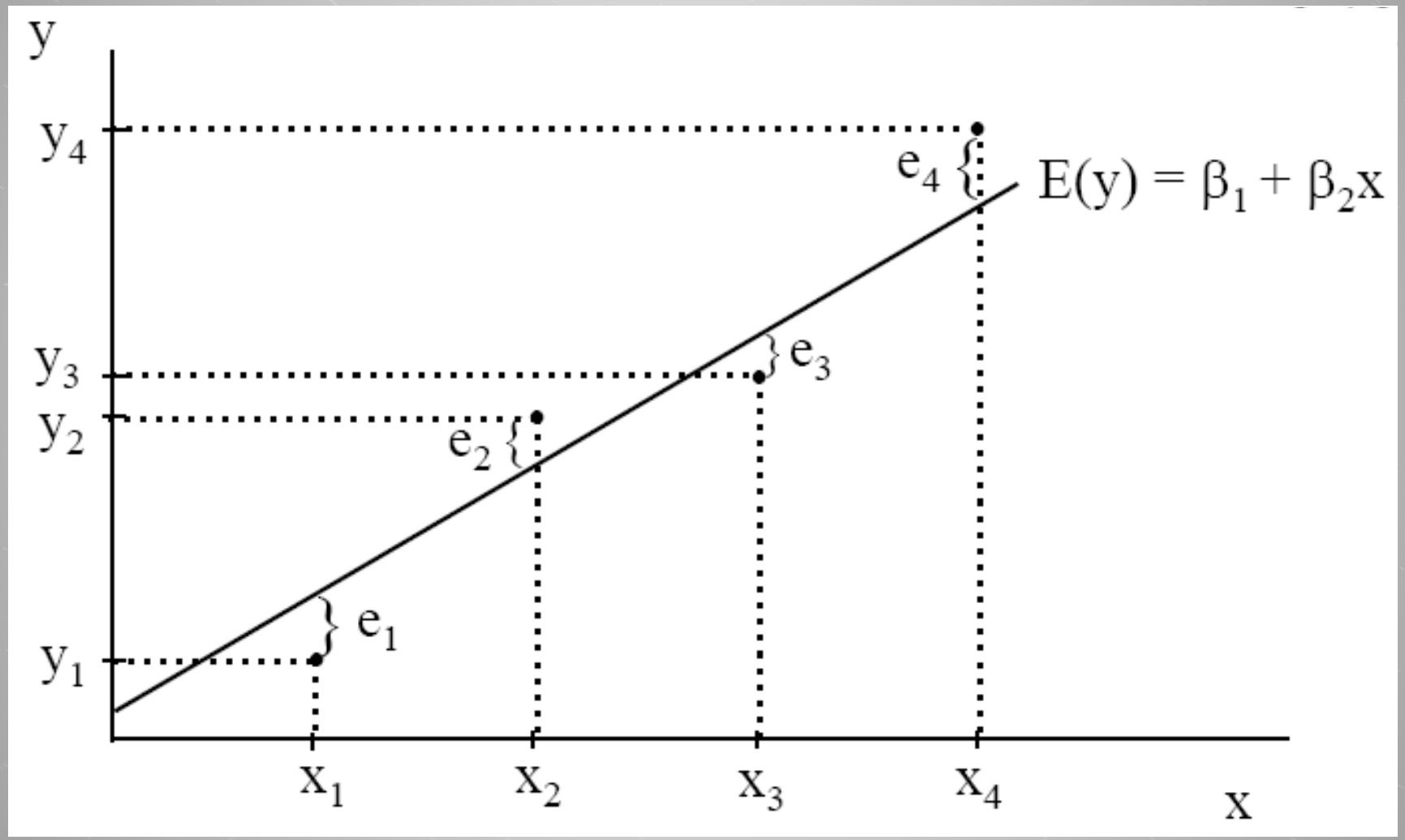
② **Random component**: $e = y - E(y) = y - \beta_1 - \beta_2 x$

This is called the **random error.**

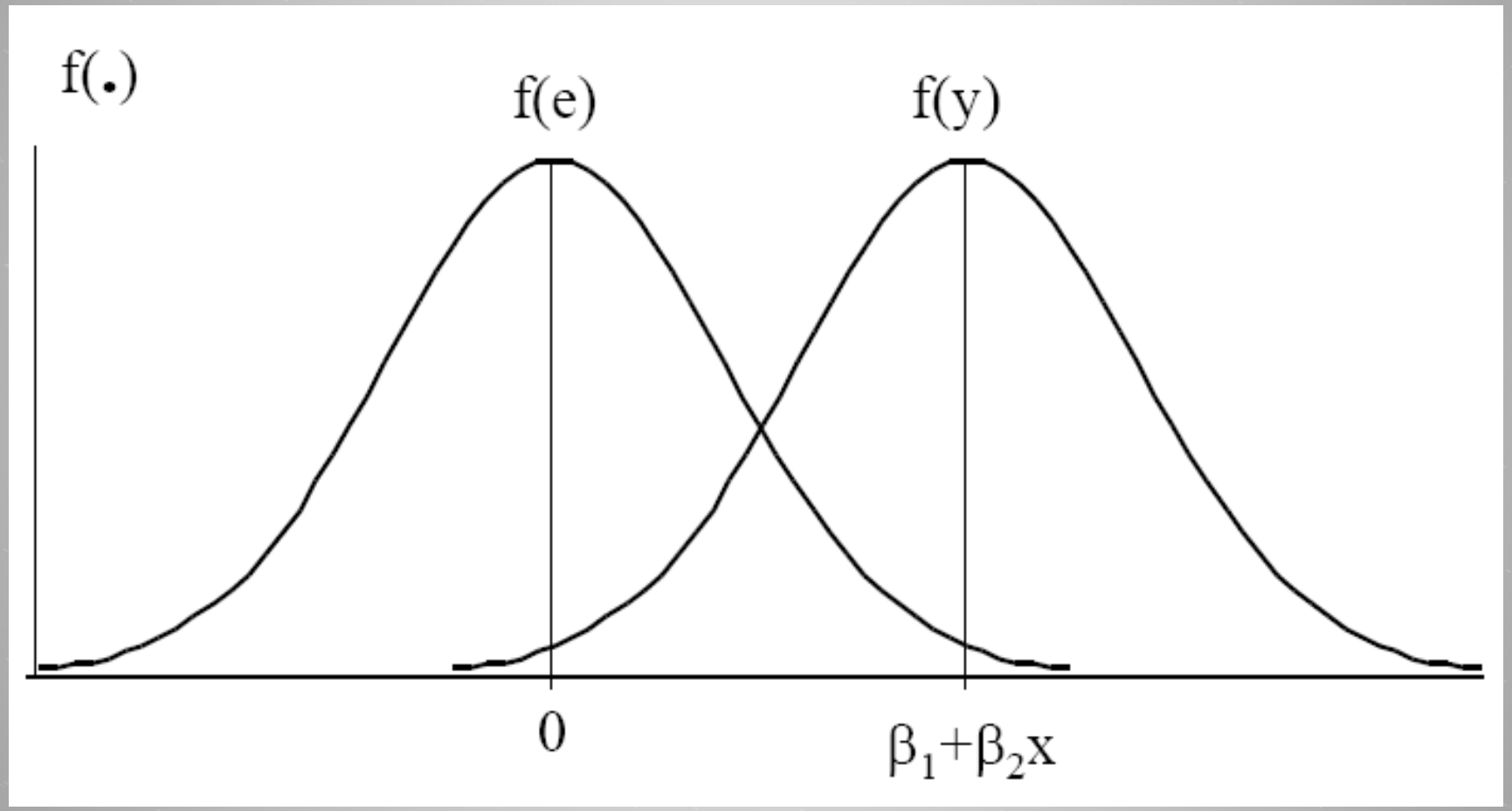
- Together E(y) and e form the **econometric model**:

$$y = \beta_1 + \beta_2 x + e$$

- y, e, 모회귀선(true regression line)의 관계



- Probability density function for e and y



<y와 e의 관계>

- 종속변수 y와 오차 e는 모두 확률변수,

  이 중 하나의 특성은 다른 하나의 특성에서 비롯됨

- y는 관찰가능, e는 관찰할 수 없음

  $e = y - (\beta_1 + \beta_2 x)$ 로 계산할 수 있을 것 같지만,

  불행히도 β_1과 β_2는 결코 알 수 없으므로

  e를 계산하는 것은 불가능함

- **Error term이 존재하는 이유**

① 설명변수의 누락

Unspecified factors / explanatory variables,
not in the model, may be in the error term.

② 부적절한 모형

Approximation error is in the error term
if relationship between y and x is not exactly
a perfectly linear relationship.

③ 경제활동의 임의성

Strictly unpredictable random behavior that
may be unique to that observation is in error.

The Error Term Assumptions

1. The value of y, for each value of x, is
$$y = \beta_1 + \beta_2 x + e$$
2. The average value of the random error e is:
$$E(e) = 0$$
3. The variance of the random error e is:
$$var(e) = \sigma^2 = var(y)$$
4. The covariance between any pair of e's is:
$$cov(e_i, e_j) = cov(y_i, y_j) = 0$$
5. x must take at least two different values so that
$x \neq c$, where c is a constant.
6. e is normally distributed with mean 0, var(e)=σ^2
(optional) $e \sim N(0, \sigma^2)$

3. 모수의 추정, Regression line의 추정

- **True (population) regression line**

$$y = \beta_1 + \beta_2 x + e$$

$$E(y) = \beta_1 + \beta_2 x$$

$$y = E(y) + e$$

- **Fitted (or estimated) regression line**

$$y = b_1 + b_2 x + \hat{e}$$

$$\hat{y} = b_1 + b_2 x$$

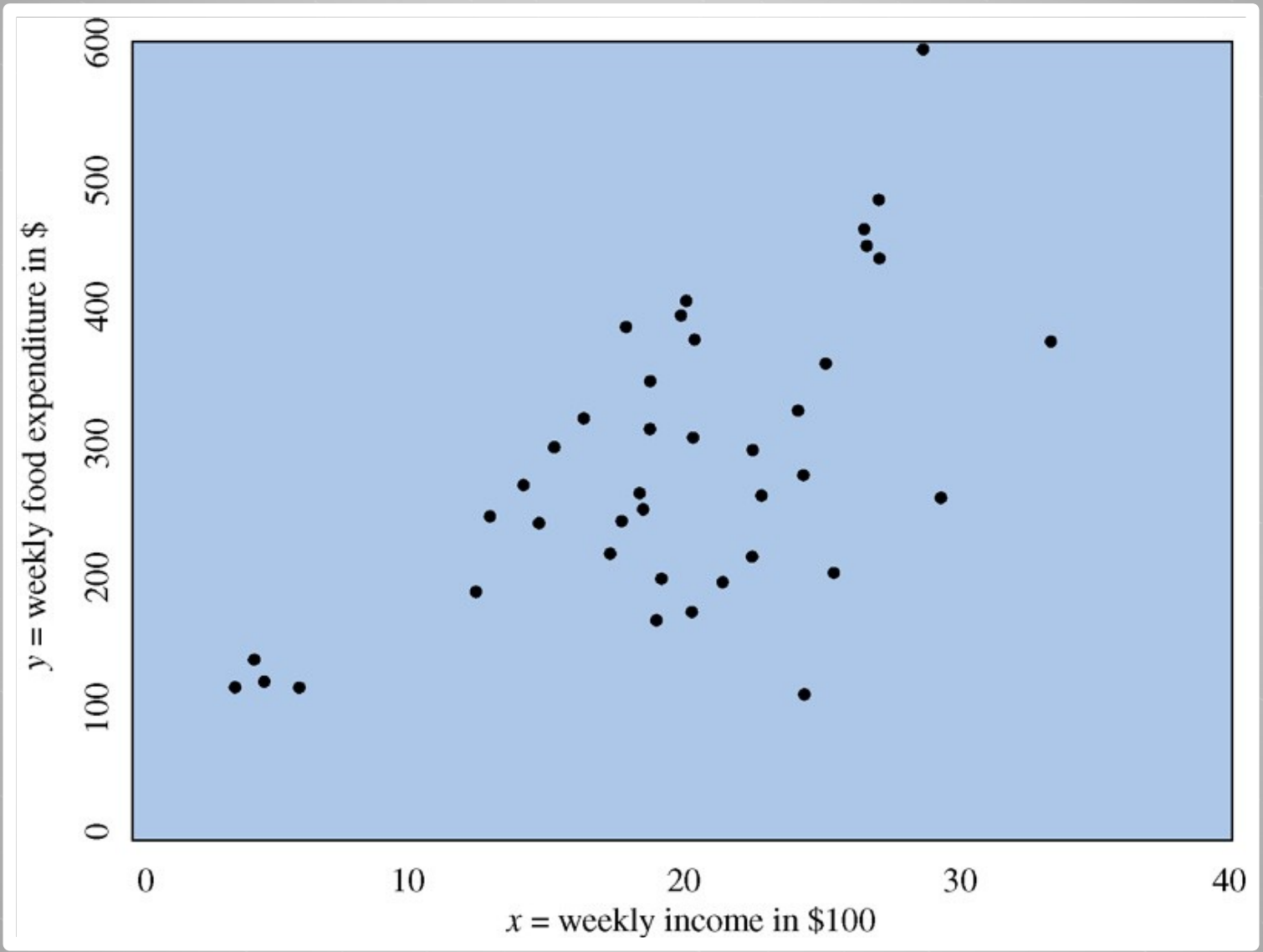
$$y = \hat{y} + \hat{e}$$

- **자료: 40 가구의 식료품 지출액과 소득** (표 <2.1>, p. 57)

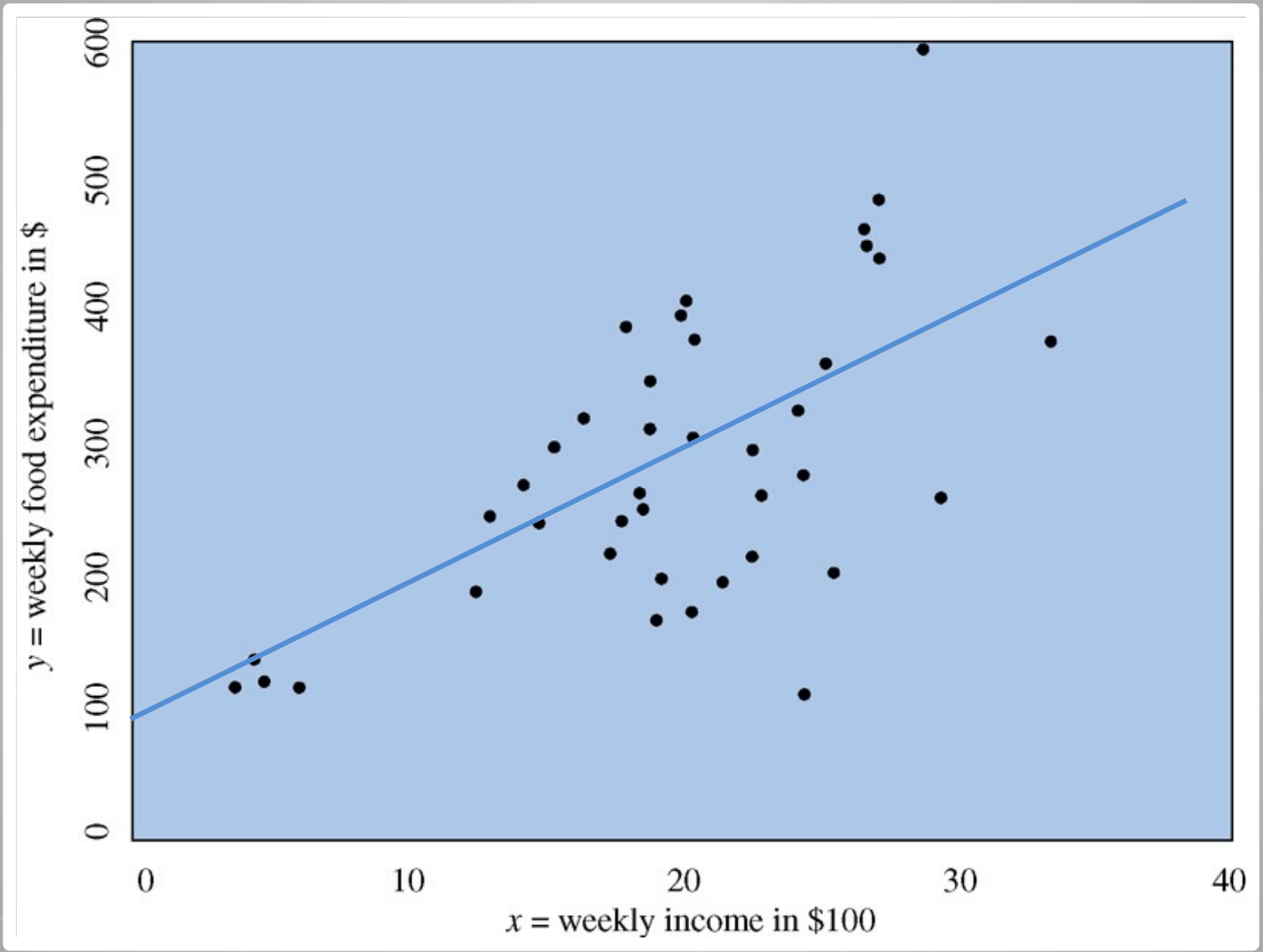
Table 2.1 **Food Expenditure and Income Data**

Observation (household)	Food expenditure ($)	Weekly income ($100)
i	y_i	x_i
1	115.22	3.69
2	135.98	4.39
	⋮	
39	257.95	29.40
40	375.73	33.40
	Summary statistics	
Sample mean	283.5735	19.6048
Median	264.4800	20.0300
Maximum	587.6600	33.4000
Minimum	109.7100	3.6900
Std. Dev.	112.7652	6.8478

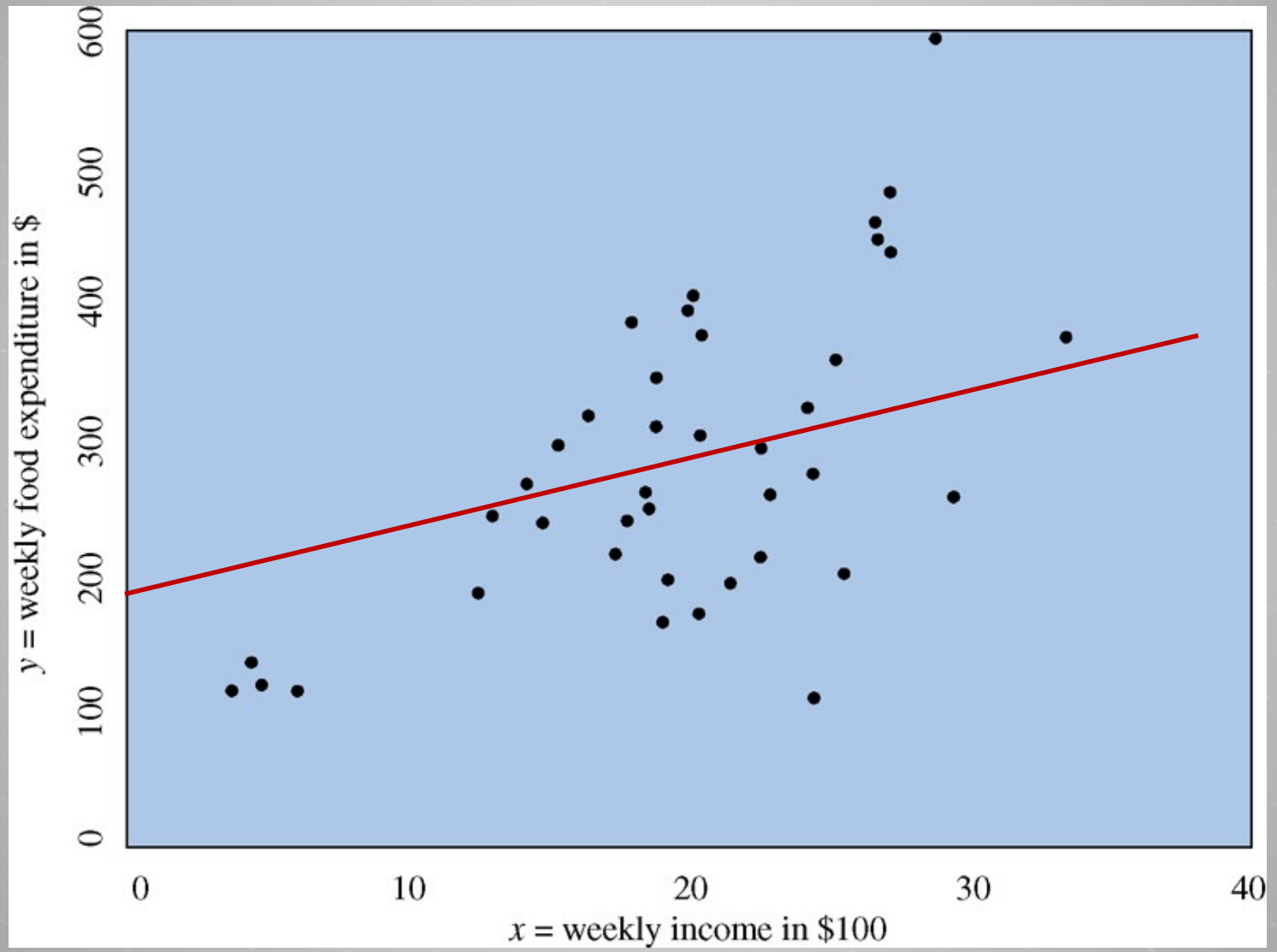
<Plot of sample data>



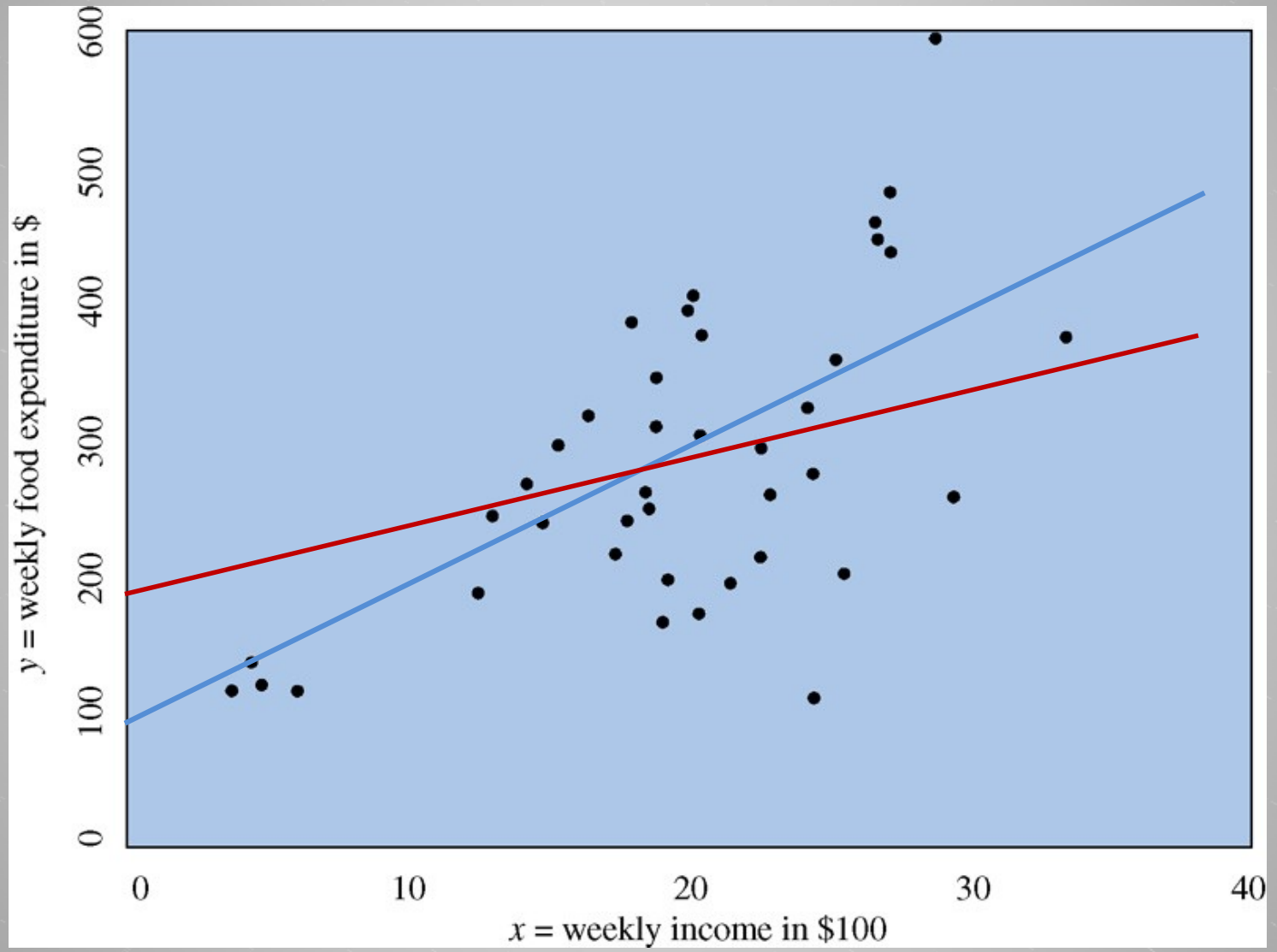
<Plot of sample data>



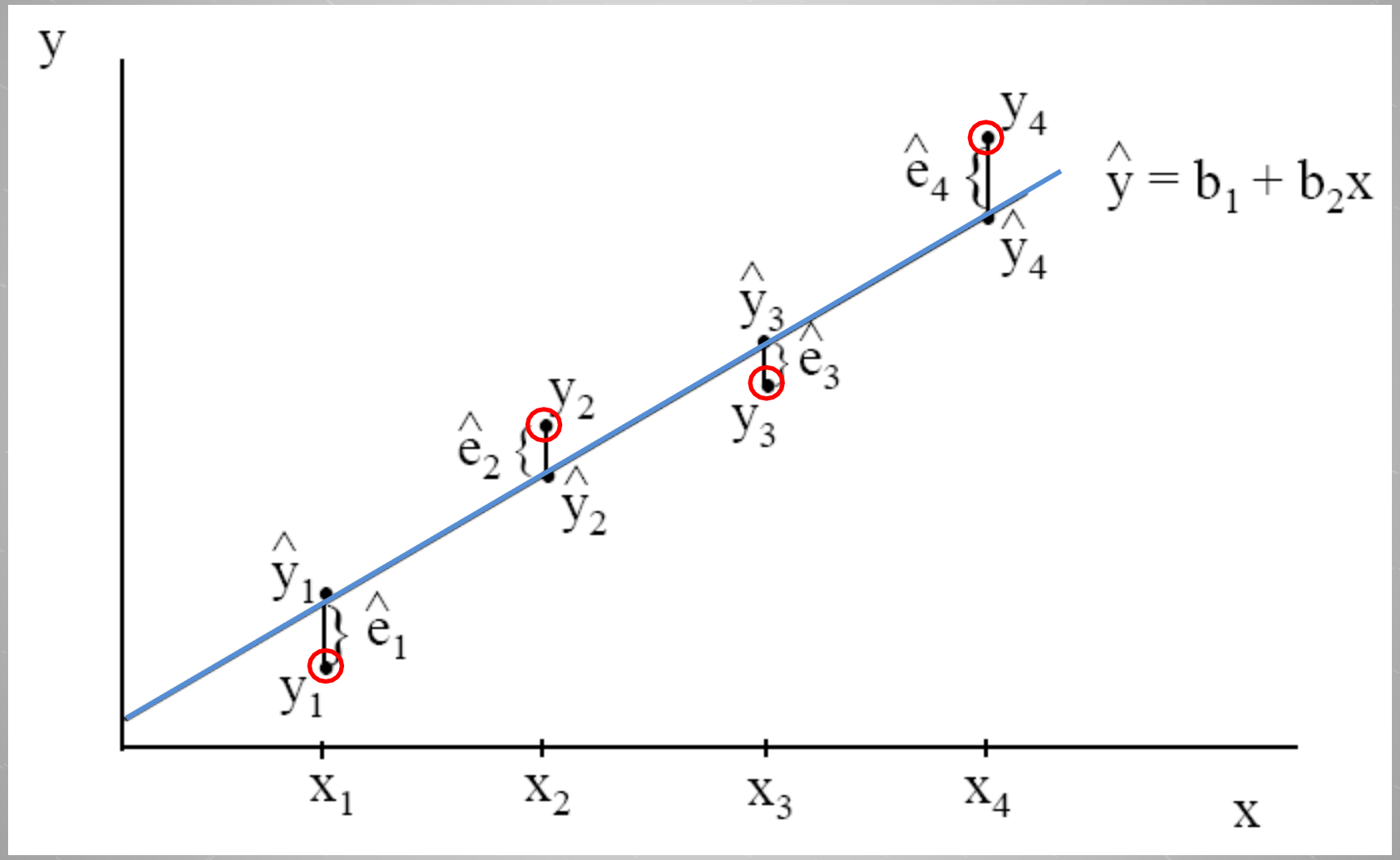
<Plot of sample data>



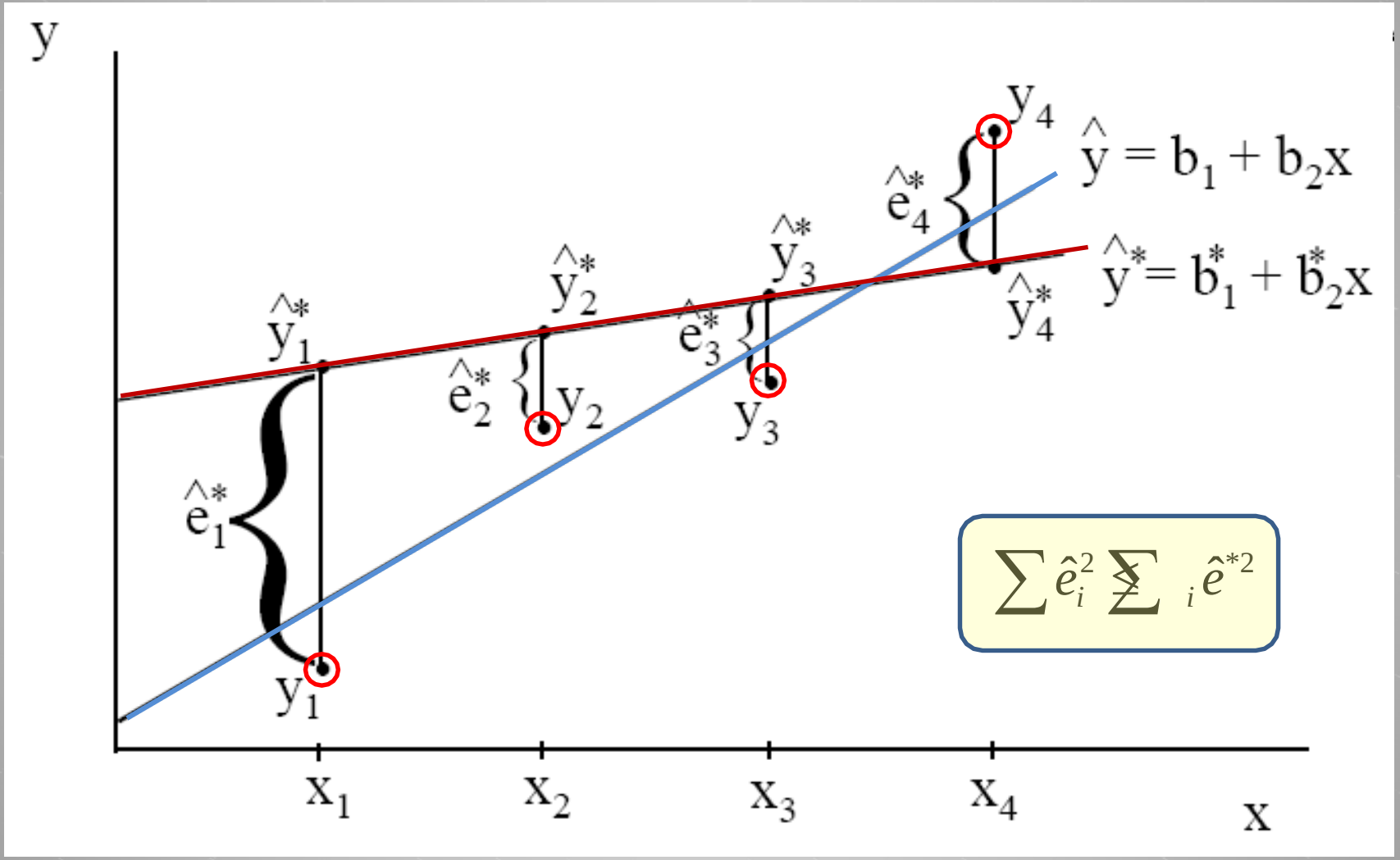
<Plot of sample data>



- relationship among y, e and the fitted regression line



- The sum of squared residuals from any other line will be larger.



3.1 최소제곱원칙 (least squares principle)

- $\sum e_i^2$ 를 최소화하는 기준으로 b_1 및 b_2를 구하여 표본회귀선을 찾으려고 함

- 이렇게 미지의 모수 β_1 및 β_2 를 구하는 방법을 최소제곱법 (ordinary least squares method: OLS)라고 함

$$y_t = \beta_1 + \beta_2 x_t + e_t$$

$$e_t = y_t - \beta_1 - \beta_2 x_t$$

Minimize error sum of squared deviations:

$$S(\beta_1,\beta_2) = \sum_{t=1}^{T}(y_t - \beta_1 - \beta_2 x_t)^2$$

$$= \sum e_t^2$$

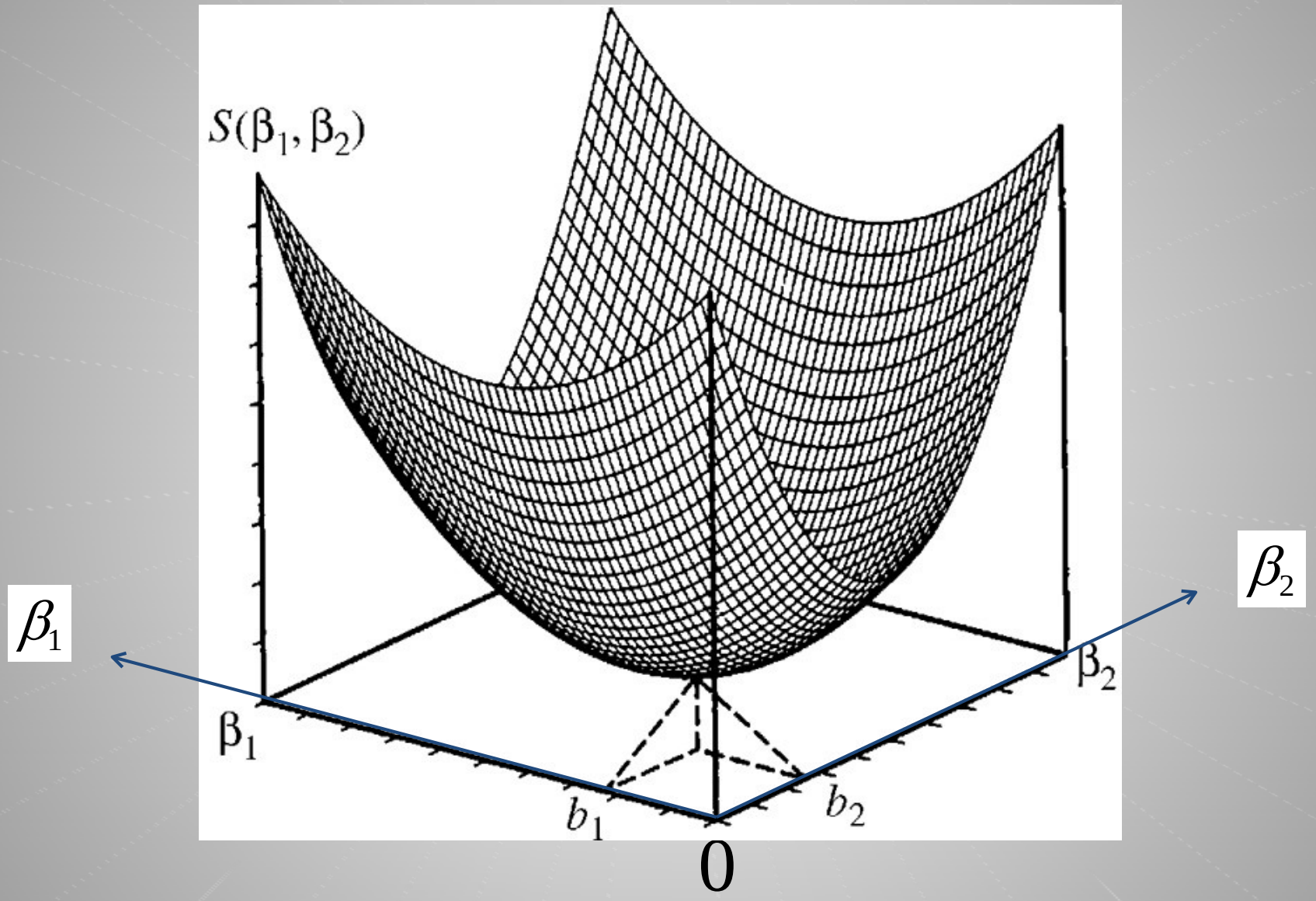
Minimize w.r.t. β_1 and β_2:

$$S(\beta_1,\beta_2) = \sum_{t=1}^{T}(y_t - \beta_1 - \beta_2 x_t)^2$$

$$\frac{\partial S(.)}{\partial \beta_1} = -2 \sum (y_t - \beta_1 - \beta_2 x_t)$$

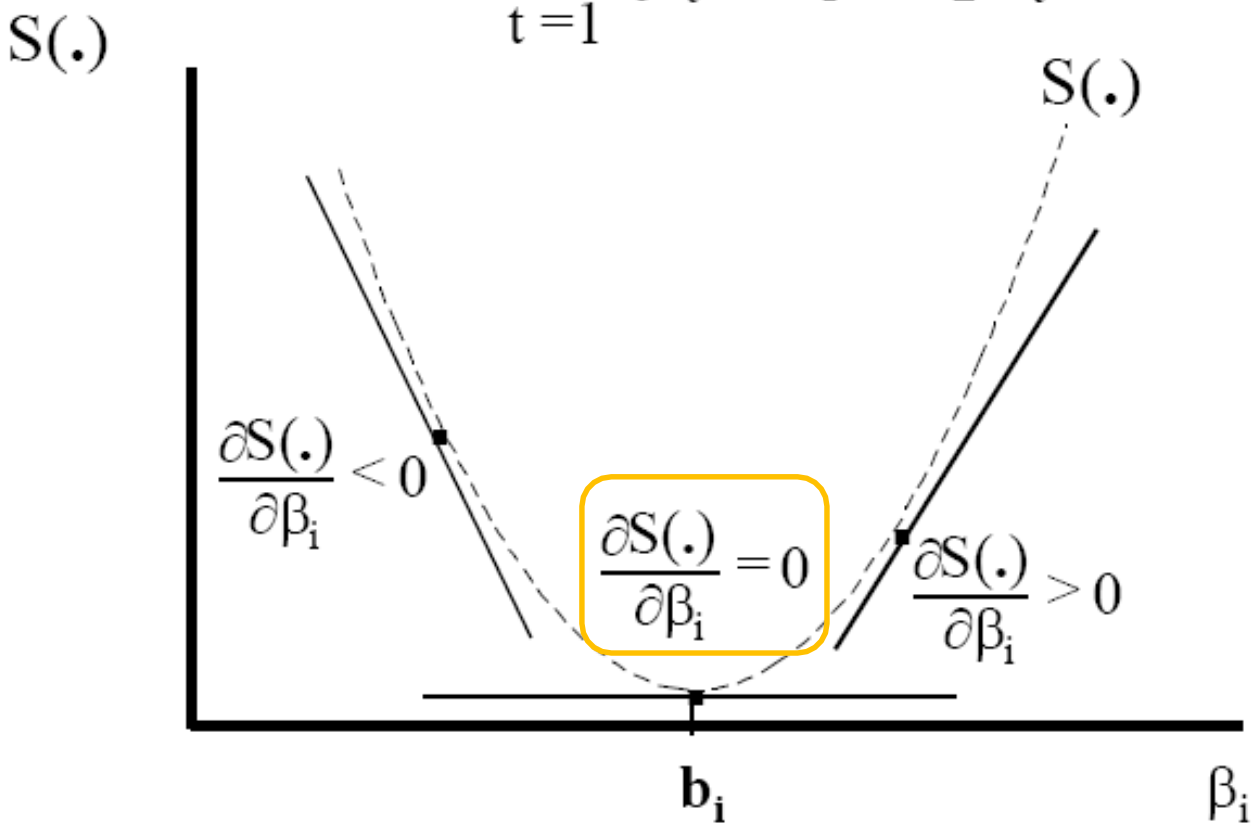
$$\frac{\partial S(.)}{\partial \beta_2} = -2 \sum x_t (y_t - \beta_1 - \beta_2 x_t)$$

Set each of these two derivatives equal to zero and solve these two equations for the two unknowns: β_1 β_2


$S(\beta_1, \beta_2)$
β_1
β_2
β_1
β_2
b_1
b_2
0

Minimize w.r.t. β_1 and β_2:

$$S(.) = \sum_{t=1}^{T} (y_t - \beta_1 - \beta_2 x_t)^2$$



To minimize S(.), you set the two derivatives equal to zero to get:

$$\frac{\partial S(.)}{\partial \beta_1} = -2 \sum (y_t - b_1 - b_2 x_t) = 0$$

$$\frac{\partial S(.)}{\partial \beta_2} = -2 \sum x_t (y_t - b_1 - b_2 x_t) = 0$$

When these two terms are set to zero, β_1 and β_2 become b_1 and b_2 because they no longer represent just any value of β_1 and β_2 but the special values that correspond to the minimum of S(.) .

$$-2\sum (y_t - b_1 - b_2 x_t) = 0$$

$$-2\sum x_t (y_t - b_1 - b_2 x_t) = 0$$

$$\sum y_t - Tb_1 - b_2 \sum x_t = 0$$

$$\sum x_t y_t - b_1 \sum x_t - b_2 \sum x_t^2 = 0$$

$$Tb_1 + b_2 \sum x_t = \sum y_t$$

$$b_1 \sum x_t + b_2 \sum x_t^2 = \sum x_t y_t$$

$$T b_1 + b_2 \sum x_t = \sum y_t$$

$$b_1 \sum x_t + b_2 \sum x_t^2 = \sum x_t y_t$$

Solve for b_1 and b_2 using definitions of $\bar{x}$ and $\bar{y}$

$$b_2 = \frac{T \sum x_t y_t - \sum x_t \sum y_t}{T \sum x_t^2 - (\sum x_t)^2}$$

$$b_1 = \bar{y} - b_2 \bar{x}$$

- **최소제곱추정량 (least squares estimator)**

  = OLS 추정량

$$b_2 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \qquad (2.7)$$

$$b_1 = \bar{y} - b_2\bar{x} \qquad (2.8)$$

3.2 식료품 지출액 함수의 추정

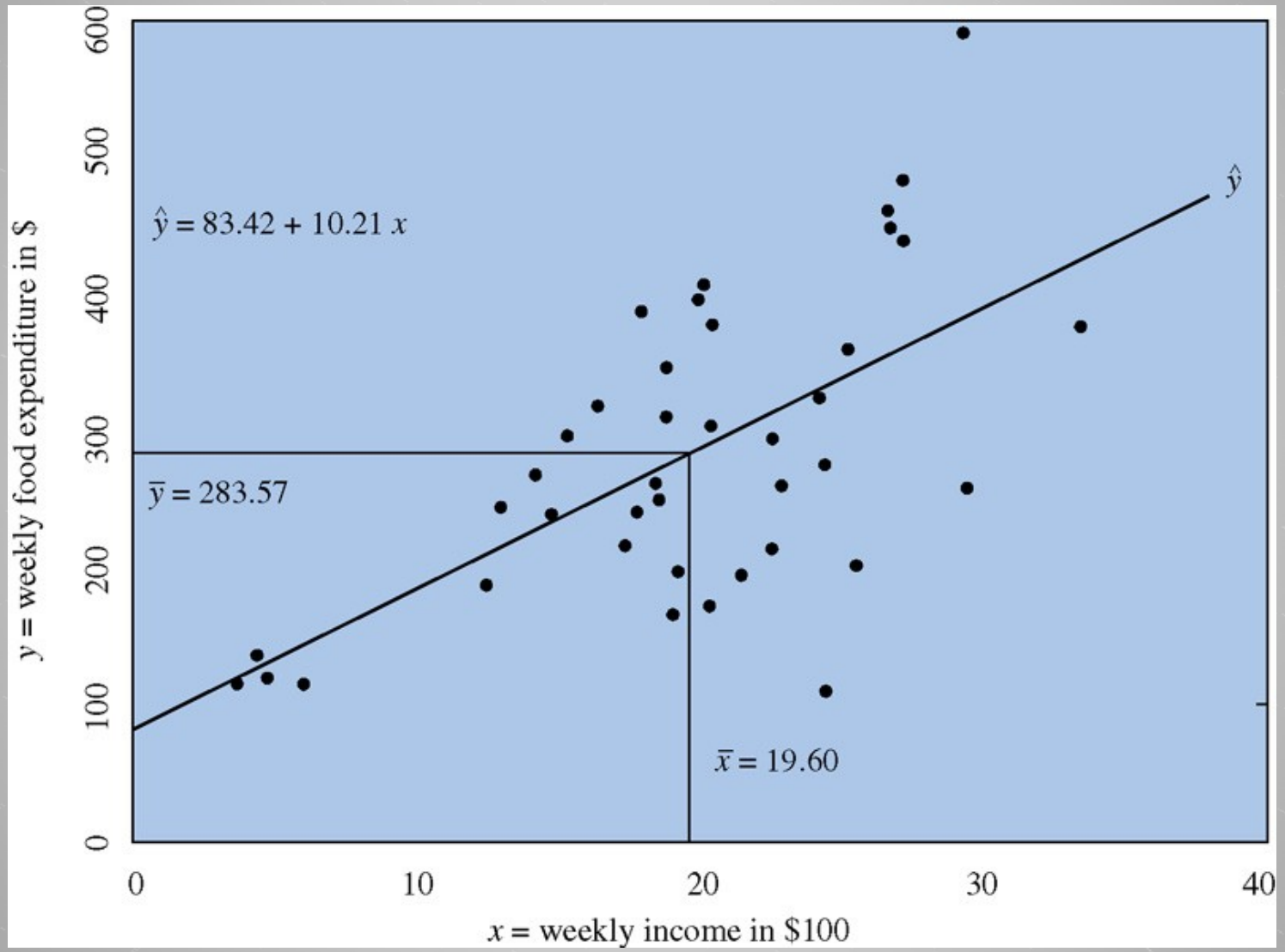
$$b_2 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{18671.2684}{1828.7876} = 10.2096$$

$$b_1 = \bar{y} - b_2\bar{x} = 283.5735 - (10.2096)(19.6048) = 83.4160$$

- *estimated* or *fitted* regression line:

$$\hat{y}_i = 83.42 + 10.21x_i$$

<estimated regression line>



3.3 추정치의 해석

$$\hat{y}_i = 83.42 + 10.21 x_i$$

$b_2 = 10.21$

⇒ 주당 소득이 1단위($100) 증가하면 식료품에 대한 주당 지출액은 10.21단위($10.21) 증가할 것으로 추정

$b_1 = 83.42$

⇒ 주당 소득이 없는 가구는 식료품에 대해 주당 83.42단위($83.42) 지출한다는 의미

✓ 상수항(절편)의 추정치를 해석할 때는 주의가 필요, $x = 0$ 근처의 값이 거의 없는 경우가 있기 때문

<탄력성의 계산>

- **활용: 식료품 지출의 소득탄력성 계산**

$$\varepsilon = \frac{\text{percentage change in } y}{\text{percentage change in } x} = \frac{\Delta y / y}{\Delta x / x} = \frac{\Delta y}{\Delta x}\frac{x}{y}$$

$$\beta_2 = \frac{\Delta E(y)}{\Delta x}$$

$$\varepsilon = \frac{\Delta E(y)/E(y)}{\Delta x/x} = \frac{\Delta E(y)}{\Delta x}\cdot\frac{x}{E(y)} = \beta_2 \cdot \frac{x}{E(y)}$$

- the elasticity at the "point of the means"

$$(\bar{x},\ \bar{y}) = (19.60, 283.57)$$

$$\hat{\varepsilon} = b_2 \frac{\bar{x}}{\bar{y}} = 10.21 \times \frac{19.60}{283.57} = 0.71$$

- 주당 가계소득이 1% 증가하면, 식료품에 대한 주당 지출은 약 0.7% 증가한다는 것을 알 수 있음

- $\hat{\varepsilon} < 1$

⇒ 식료품은 '사치품'이 아닌 **'필수품'**으로 분류될 수 있음

<예측>

- 추정된 식은 예측에 활용될 수 있음
- 주당 소득이 $2,000인 가계의 식료품 지출액은 얼마일까?

$$\hat{y}_i = 83.42 + 10.21 x_i = 83.42 + 10.21(20) = 287.61$$

- 주당 소득이 $2,000인 가계는 식료품에 주당 $287.61을 지출한다고 **예측할 수 있음**

▪ 계량경제 SW를 이용한 추정 결과

<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

▪ 비선형모형을 선형모형으로 전환

- 선형회귀모형에서 '선형'의 용어는 '모수'에 대해 선형이라는 것이고, '변수'에 대해서는 비선형이어도 됨

- 다음과 같은 모형도 선형회귀모형으로 변형될 수 있음
- log-log model: $\log(y) = \beta_1 + \beta_2 \ln(x)$
- semi-log model: $y = \beta_1 + \beta_2 \ln(x)$
- $y = \beta_1 + \beta_2 \frac{1}{x}$
- $y = \beta_1 + \beta_2 x_3$

$$YY = \log(y), \quad XX = \log(x)$$

$$YY = \beta_1 + \beta_2 XX$$

4. 최소제곱 추정량에 대한 평가

- **OLS 추정량은 확률변수**

- <표 2.1>의 표본에서 계산한 OLS 추정치는

$$b_1 = 83.42, \quad b_2 = 10.21$$

- b_1, b_2 값은 표본을 뽑을 때마다 다른 값으로 계산됨

  $\Rightarrow$ b_1, b_2 는 확률변수

▪ OLS 추정량은 확률변수

- <표 2.1>과 다른 표본 10개를 조사하여 구한 b_1, b_2 추정치

Table 2.2 **Estimates from 10 Samples**

Sample	b_1	b_2
1	131.69	6.48
2	57.25	10.88
3	103.91	8.14
4	46.50	11.90
5	84.23	9.29
6	26.63	13.55
7	64.21	10.93
8	79.66	9.76
9	97.30	8.05
10	95.96	7.77

- 표본마다 다른 값으로 추정됨, 즉 확률변수임

▪ OLS 추정량은 확률변수

- b_1, b_2 추정량은 확률변수
- b_1, b_2 추정량의 평균, 분산, 공분산, 확률분포는 무엇인가?
- β_1, β_2 에 근접한 추정치를 얻을 수 있는 확률이 더 높은 다른 추정량이 있는가?

- b_1, b_2 are **unbiased estimators** of β_1, β_2.

$$E(b_1) = \beta_1 \qquad E(b_2) = \beta_2$$

- **불편추정량의 의미**
  - b_1, b_2의 확률분포의 평균은 β_1, β_2
  - b_1, b_2는 β_1, β_2에 근접한 값을 구해 줄 가능성이 높음
- 그러나 모형설정이 잘못되면 (중요한 설명변수 누락되면)

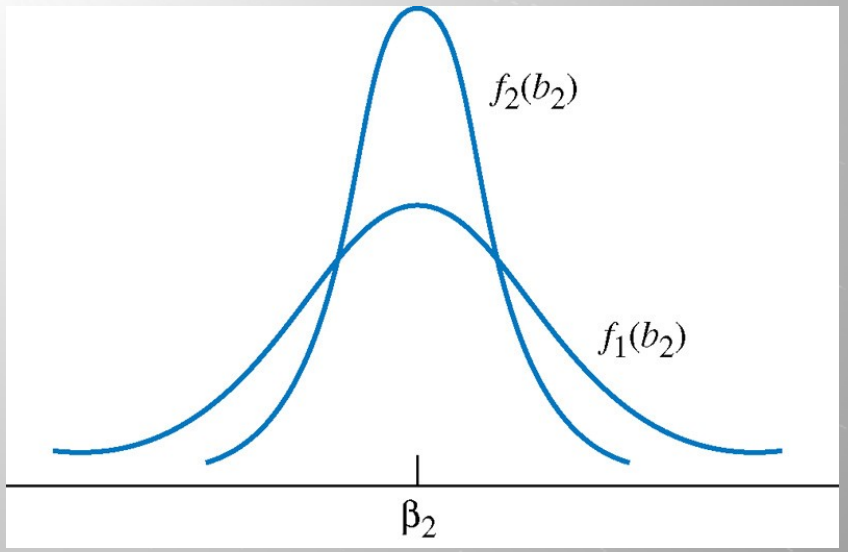
  ⇒ 불편추정량 안됨

$$E(b_1) \neq \beta_1 \qquad E(b_2) \neq \beta_2$$

■ b_1, b_2 의 **분산**

$$\text{var}(b_2) = E\left[b_2 - E(b_2)\right]^2$$

- b_1, b_2 는 얼마나 정확한 추정량인가?
- 분산이 적을수록 추정량의 정확도가 높아짐
- β_1, β_2 에 근접한 추정량 b_1, b_2 을 얻는 것이 목표임
- $f_2(b_2)$ 가 더 좋은 추정량



▪ OLS 추정량의 분산과 공분산 추정식

$$\mathrm{var}(b_1) = \sigma^2 \left[\frac{\sum x_i^2}{N \sum (x_i - x)^2} \right] \qquad (2.14)$$

$$\mathrm{var}(b_2) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \qquad (2.15)$$

$$\mathrm{cov}(b_1, b_2) = \sigma^2 \left[\frac{-\bar{x}}{\sum (x_i - \bar{x})^2} \right] \qquad (2.16)$$

▪ 분산과 공분산에 영향을 미치는 요인들

① σ^2 이 작을수록 b_1, b_2 의 분산이 작아짐

- y 값이 $E(y)$ 에 집중되어 있을수록 σ^2 작아짐
- 확률분포 $f(Y|X)$ 의 퍼진 정도가 작을수록

  추정량 b_1, b_2 의 정확도가 높아짐

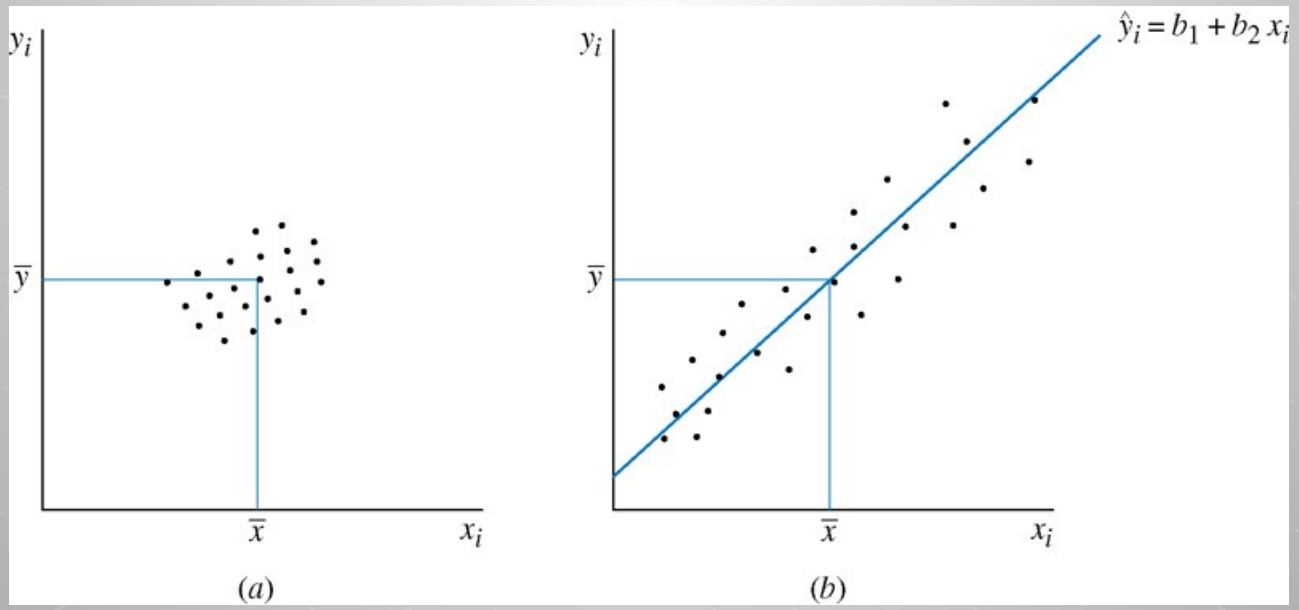
$$\text{var}(b_1) = \sigma^2 \left[\frac{\sum x_i^2}{N \sum (x_i - \bar{x})^2} \right]$$

$$\text{var}(b_2) = \frac{\sigma^2}{\sum (x_i - x)^2}$$

▪ 분산과 공분산에 영향을 미치는 요인들 (계속)

② $\sum(X_t - \bar{X})^2$ 이 클수록 b_1, b_2 의 분산이 작아짐

- 설명변수 X 의 값이 평균에서 퍼져 있을수록 추정량 b_1, b_2 의 정확도가 높아짐



▪ 분산과 공분산에 영향을 미치는 요인들 (계속)

③ 표본크기 N 이 클수록 b_1, b_2 의 분산이 작아짐

- 표본을 많이 수집할수록 모수 β_1, β_2 를 추정하기 위한 추정량 b_1, b_2 의 정확도가 높아짐

Cf. $Var(b_2)$, $Cov(b_1, b_2)$ 계산식 분모의 $\sum (X_t - \bar{X})^2 \geq 0$

따라서 표본이 많을수록 분모는 커짐에 유의

$$\text{var}(b_1) = \sigma^2 \left[\frac{\sum x_i^2}{N \sum (x_i - \bar{x})^2} \right]$$

$$\text{var}(b_2) = \frac{\sigma^2}{\sum (x_i - x)^2}$$

▪ 분산과 공분산에 영향을 미치는 요인들 (계속)

④ X 의 표본값들이 0에서 멀리 떨어져 있을수록

$\Rightarrow$ b_1 의 분산이 커짐

$\Rightarrow$ 모수 β_1 을 정확하게 추정하기가 어려워짐

$\Rightarrow$ b_1 추정치를 해석하는 것이 어려워짐

$$\text{var}(b_1) = \sigma^2 \left[\frac{\sum x_i^2}{N \sum (x_i - x)^2} \right]$$

Cf. $Var(b_1)$ 계산식 분자에 $\sum X^2$ 이 있음

X 의 표본값들이 0에서 멀리 떨어져 있을수록

$\sum X^2$은 큰 값이 되고, b_1 의 분산이 커짐에 유의

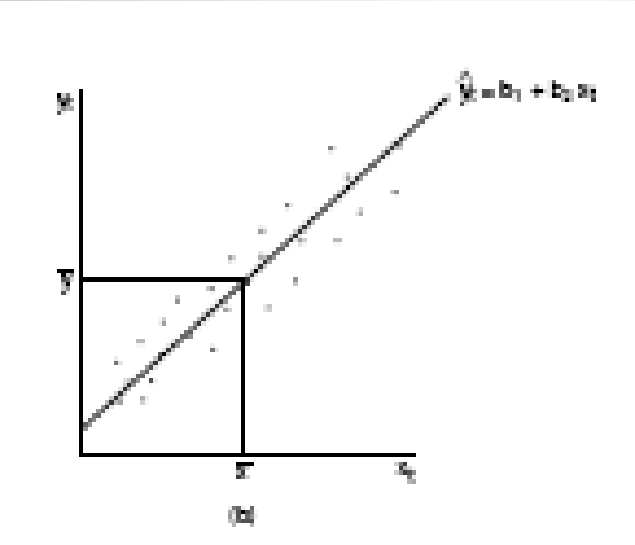
▪ 분산과 공분산에 영향을 미치는 요인들 (계속)

⑤ $Cov(b_1, b_2) < 0$

⇒ b_1 과 b_2 값은 반대 방향 ($\bar{X} > 0$ 인 경우)

(즉, 하나가 증가하면 다른 하나는 감소)

Cf. 추정된 회귀선은 $(\bar{X}, \bar{Y})$ 을 반드시 통과함에 유의



$$\text{cov}(b_1, b_2) = \sigma^2 \left[\frac{-\bar{x}}{\sum (x_i - \bar{x})^2} \right]$$

- b_1, b_2는 **선형추정량**

- b_2는 Y_t의 가중합산, 즉 선형조합(linear combination)

  $b_2 = \sum w_t Y_t$, w_t 는 상수 / 교과서 식 (2.10) 참조

- b_1 도 Y_t의 선형조합

  Cf. $b_1 = \bar{Y} - b_2 \bar{X}$

➢ b_1, b_2 는 β_1, β_2의 **선형 불편추정량**

5. 가우스-마코프(Gauss-Markov) 정리

- 단순회귀모형에 관한 앞의 가정(SR1-SR5) 하에서

  최소제곱추정량 b_1, b_2 는 β_1, β_2 의

  모든 선형 불편 추정량 중에서 **최소의 분산**을 가진다.

- b_1, b_2 는 β_1, β_2 의 **최우수 선형 불편 추정량**이다. (Best Linear Unbiased Estimators: BLUE)

- OLS estimators are **BLUE**.

- Cf. 증명 / 교과서 pp. 91-92.

- 가우스-마코프 정리는 정규성 가정(SR6)에 의존 않음

▪ OLS 추정의 기본가정

Assumptions of the Simple Linear Regression Model

SR1. $y_t = \beta_1 + \beta_2 x_t + e_t$

SR2. $E(e_t) = 0 \Leftrightarrow E(y_t) = \beta_1 + \beta_2 x_t$

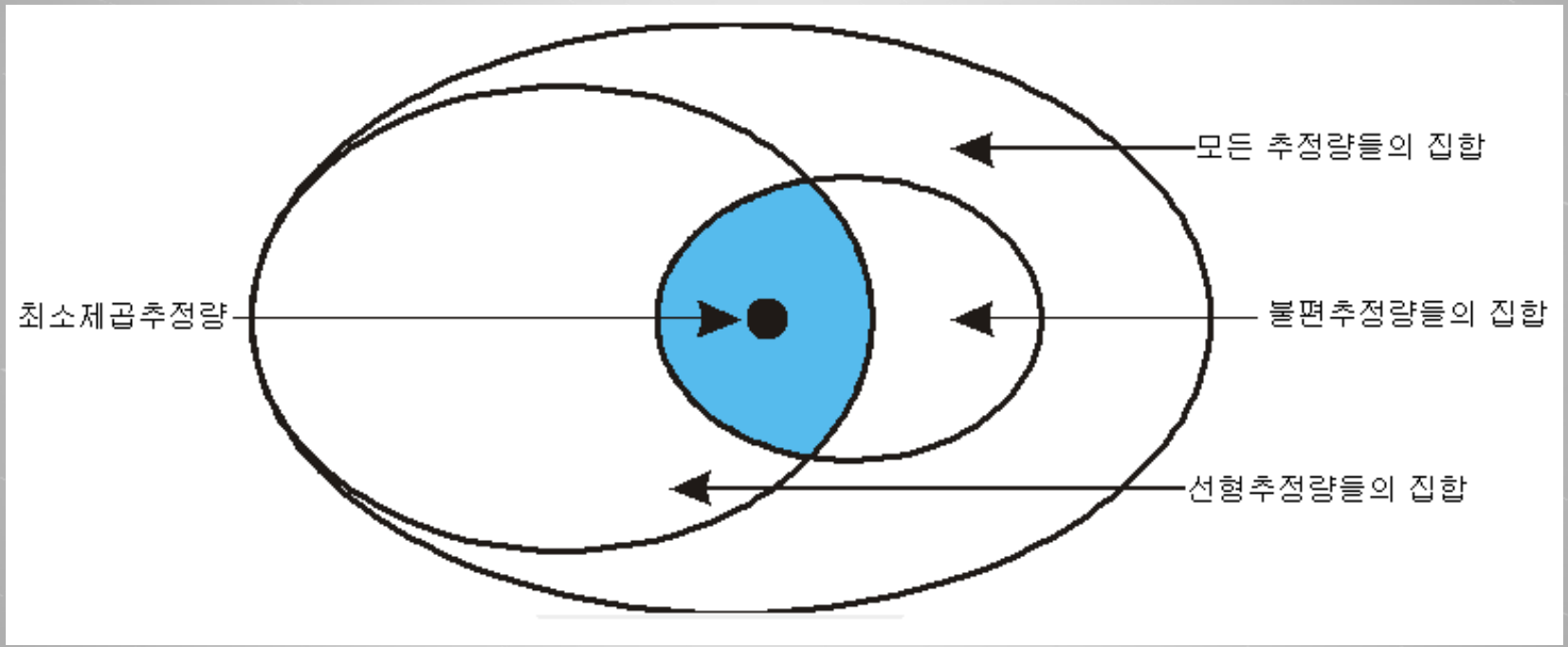
SR3. $\text{var}(e_t) = \sigma^2 = \text{var}(y_t)$

SR4. $\text{cov}(e_i, e_j) = \text{cov}(y_i, y_j) = 0$

SR5. x_t is not random and takes at least two values

SR6. $e_t \sim N(0, \sigma^2) \Leftrightarrow y_t \sim N[(\beta_1 + \beta_2 x_t), \sigma^2]$ *(optional)*

- **OLS** estimators are **BLUE**.



6. OLS 추정량의 확률분포

- 정규성 가정(SR6) $e_t \sim N(0,\sigma^2)$ 이 충족되면

  OLS 추정량 b_1, b_2 는 정규분포를 함

  - 증명: e_t 가 정규분포하면 Y_t 도 정규분포함,

    b_1, b_2 는 Y_t 의 선형추정량,

    정규 확률변수의 선형결합도 정규 확률변수. 끝.

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2 \sum x_i^2}{N\sum (x_i - x)^2}\right) \qquad b_2 \sim N\left(\beta_2, \frac{\sigma^2}{\sum (x_i - x)^2}\right)$$

- **표본의 크기와 OLS 추정량의 확률분포**
- 가정 SR1-SR5가 준수되고, 표본크기가 충분히 큰 **대표본**의 경우

  ⇒ OLS 추정량은 정규분포에 근접한 분포를 한다.

  (중심극한정리, Central Limit Theorem)

- 대표본은 얼마나 커야 하나?

  ⇒ $N = 30$ 혹은 $N = 50$

- 왜 OLS 추정량의 확률분포에 관심을 가지나?

- **표본의 크기와 OLS 추정량의 일치성(consistency)**
- OLS 추정량은 일치추정량인가?
- 즉 $N \to \infty$ 일 때, $b_1 \to \beta_1$, $b_2 \to \beta_2$ 인가?
- 이렇게 되기 위해서는

  $N \to \infty$ 일 때, $Var(b_1) \to 0$, $Var(b_2) \to 0$ 이어야 함
- 표본크기가 증가할수록, 아래 분산식의 분모가 커짐

  따라서 $Var(b_1) \to 0$, $Var(b_2) \to 0$, 즉 일치추정량임

$$\text{var}(b_1) = \sigma^2 \left[\frac{\sum x_i^2}{N \sum (x_i - x)^2} \right]$$

$$\text{var}(b_2) = \frac{\sigma^2}{\sum (x_i - x)^2}$$

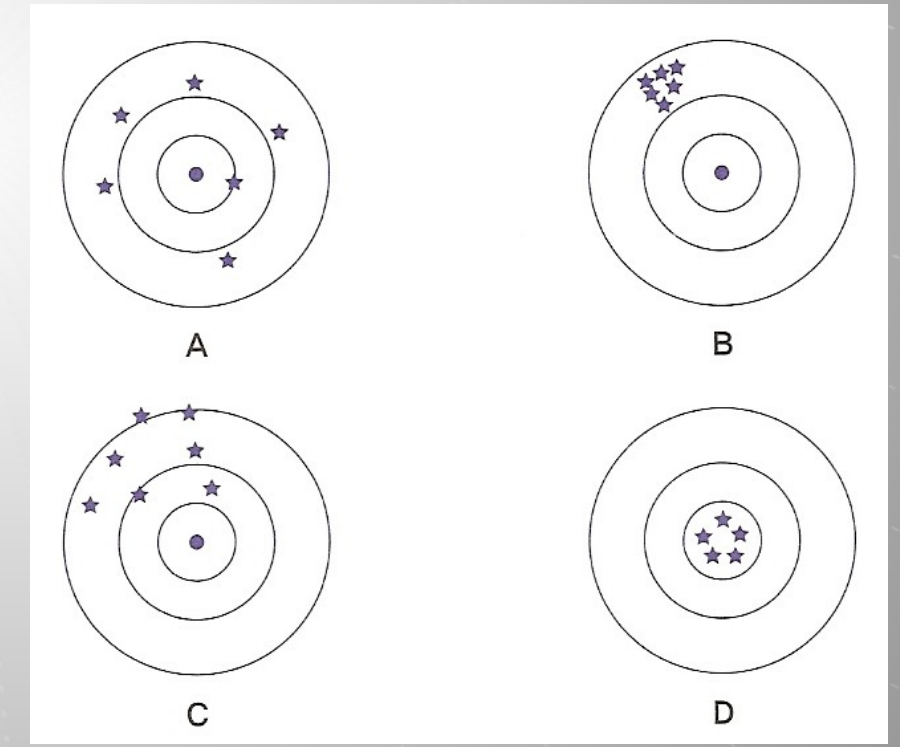
- 정확도 높은 추정치를 얻기 위해서는 $N \to \infty$ 필요

❖ 점추정량의 바람직한 성질

- 불편성: 추정량의 기대값이 모수와 일치함
- 일치성: 표본의 크기가 증가하면 추정량이 모수와 일치함
- 최소분산: 추정량의 분산이 적어 추정값의 변동폭이 적음

➢ 그림

4개의 총에 의한 사격결과



7. 오차항의 분산에 대한 추정

- 오차항의 분산 σ^2 은 추정해야 할 미지의 모수임
  - 이 값을 알아야 b_1, b_2 의 분산을 계산할 수 있음
- $Var(e_t) = \sigma^2 = E[e_t - E(e_t)]^2 = E(e_t^2)$ 이므로
  오차항의 분산은 $\hat{\sigma}^2 = \frac{\sum e_i^2}{N}$ 으로 추정할 수 있음
- e_t 는 관찰 불가능하므로, OLS 잔차로 대체시켜 봄

$$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{N}$$

2.7 오차항의 분산에 대한 추정 (계속)

$$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{N}$$

- 이 값은 편의가 있으므로, 약간 수정하면 불편추정량 됨

$$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{N-2}$$

  - 여기서 "2"는 회귀식에서의 모수(β_1, β_2)의 개수
- 오차항 분산은 위와 같이 추정함
- $\hat{\sigma}^2$ 은 σ^2의 불편추정량임, $E(\hat{\sigma}^2) = \sigma^2$

2.7.1 OLS 추정량의 분산 및 공분산 추정식

$$\widehat{\mathrm{var}}(b_1) = \hat{\sigma}^2 \left[\frac{\sum x_i^2}{N \sum (x_i - x)^2} \right]$$

$$\mathrm{se}(b_1) = \sqrt{\widehat{\mathrm{var}}(b_1)}$$

$$\widehat{\mathrm{var}}(b_2) = \frac{\hat{\sigma}^2}{\sum (x_i - x)^2}$$

$$\mathrm{se}(b_2) = \sqrt{\widehat{\mathrm{var}}(b_2)}$$

$$\widehat{\mathrm{cov}}(b_1, b_2) = \hat{\sigma}^2 \left[\frac{-\bar{x}}{\sum (x_i - \bar{x})^2} \right]$$

$$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{N - 2}$$

2.7.2 식료품 지출액 경우의 분산 및 공분산 추정

Table 2.3 **Least Squares Residuals**

x	y	$\hat{y}$	$\hat{e} = y - \hat{y}$
3.69	115.22	121.09	−5.87
4.39	135.98	128.24	7.74
4.75	119.34	131.91	−12.57
6.03	114.96	144.98	−30.02
12.47	187.05	210.73	−23.68

$$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{N-2} = \frac{304505.2}{38} = 8013.29$$

▪ 식료품 지출액 경우의 분산-공분산 행렬 추정

$$\begin{pmatrix} \widehat{\mathrm{var}}(b_1) & \widehat{\mathrm{cov}}(b_1, b_2) \\ \widehat{\mathrm{cov}}(b_1, b_2) & \widehat{\mathrm{var}}(b_2) \end{pmatrix}$$

	C	*INCOME*
C	1884.442	-85.90316
INCOME	-85.90316	4.381752

▪ 식료품 지출액 경우의 분산, 공분산, 표준오차 추정

$$\widehat{\mathrm{var}}(b_1) = 1884.442$$

$$\widehat{\mathrm{var}}(b_2) = 4.381752$$

$$\widehat{\mathrm{cov}}(b_1, b_2) = -85.90316$$

$$\mathrm{se}(b_1) = \sqrt{\widehat{\mathrm{var}}(b_1)} = 43.410$$

$$\mathrm{se}(b_2) = \sqrt{\widehat{\mathrm{var}}(b_2)} = 2.093$$

- **계량경제 SW를 이용한 추정 결과**

<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

<유용한 동영상> (강의 보충)

단순회귀분석의 실습 (엑셀)
https://www.youtube.com/watch?v=qbIVZM0c5Qg&t=1s

다중회귀분석의 실습 (엑셀)
https://www.youtube.com/watch?v=8qx5x-4h13U

<과제> (교과서 연습문제 풀이)

2.1
2.2
2.4
2.7
2.11

※ 참고: 필요한 data는 WILEY 교과서 홈페이지에 있음

- http://principlesofeconometrics.com/poe3/poe3.htm