Principles of Econometrics (3e)

Ch. 8 이분산

2013년 1학기

윤성민

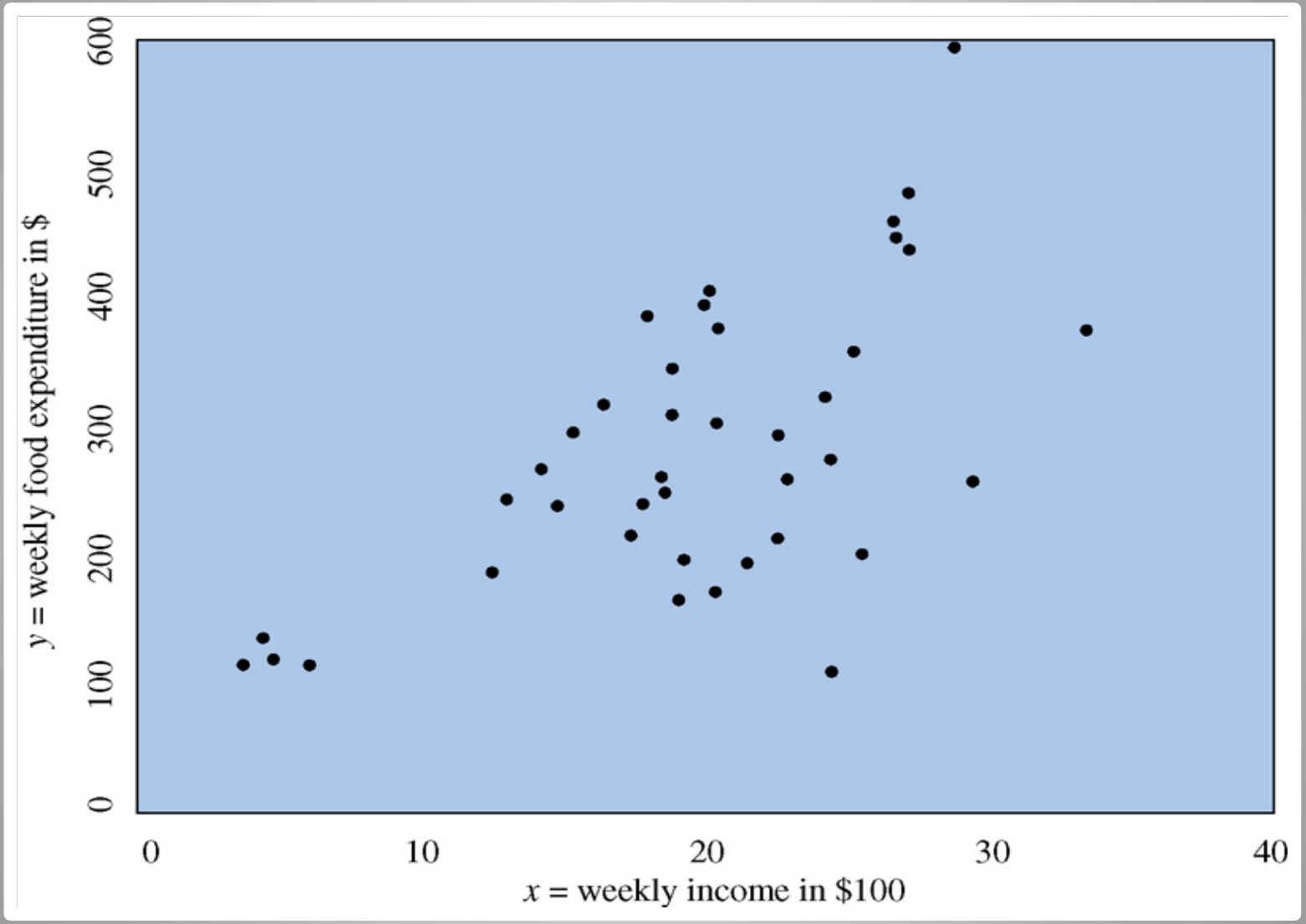
8.1. 이분산의 본질

(예) 식료품 지출 / 식료품 지출과 소득에 관한 40개 표본

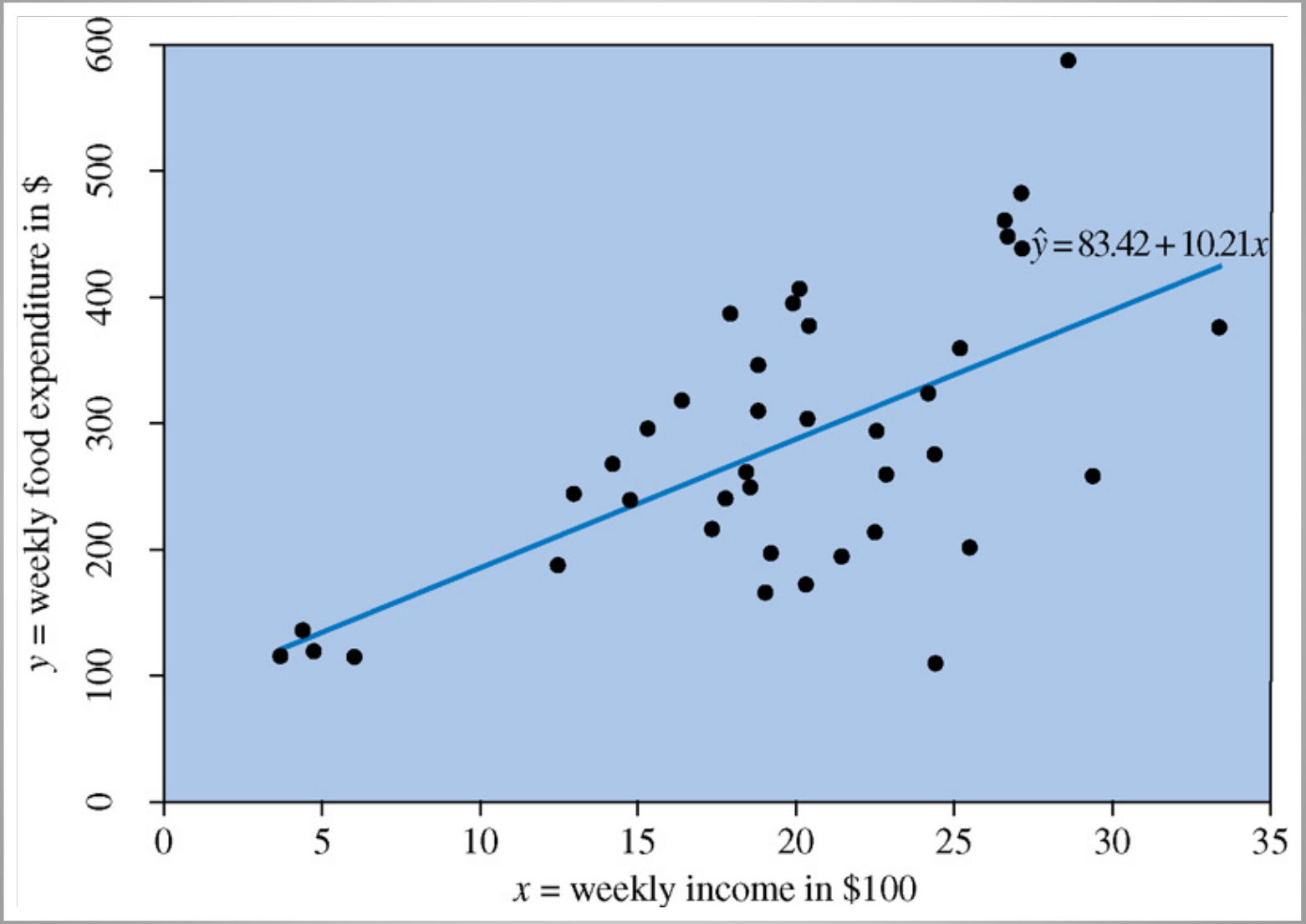
Table 2.1 Food Expenditure and Income Data

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

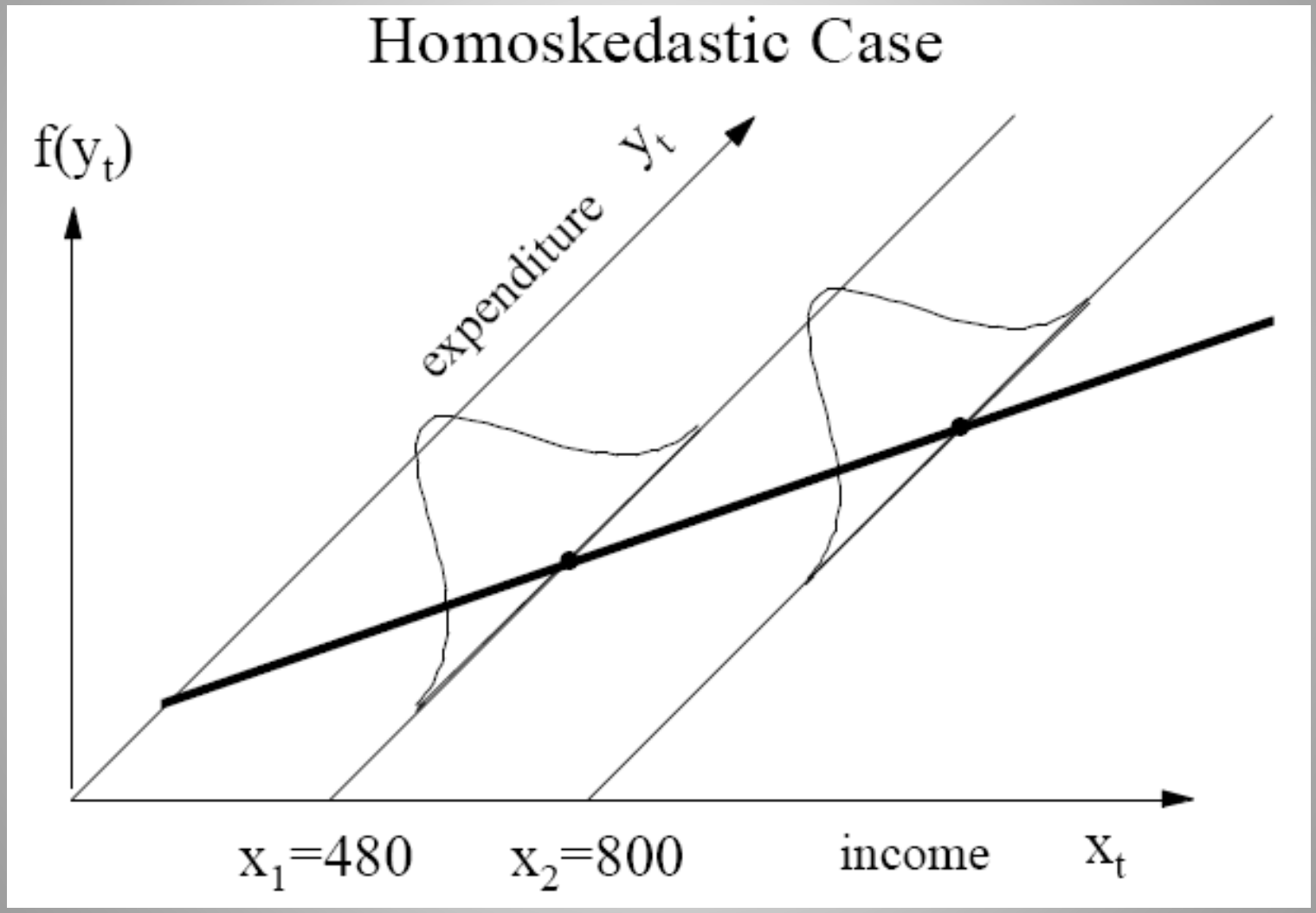
<Plot of sample data>



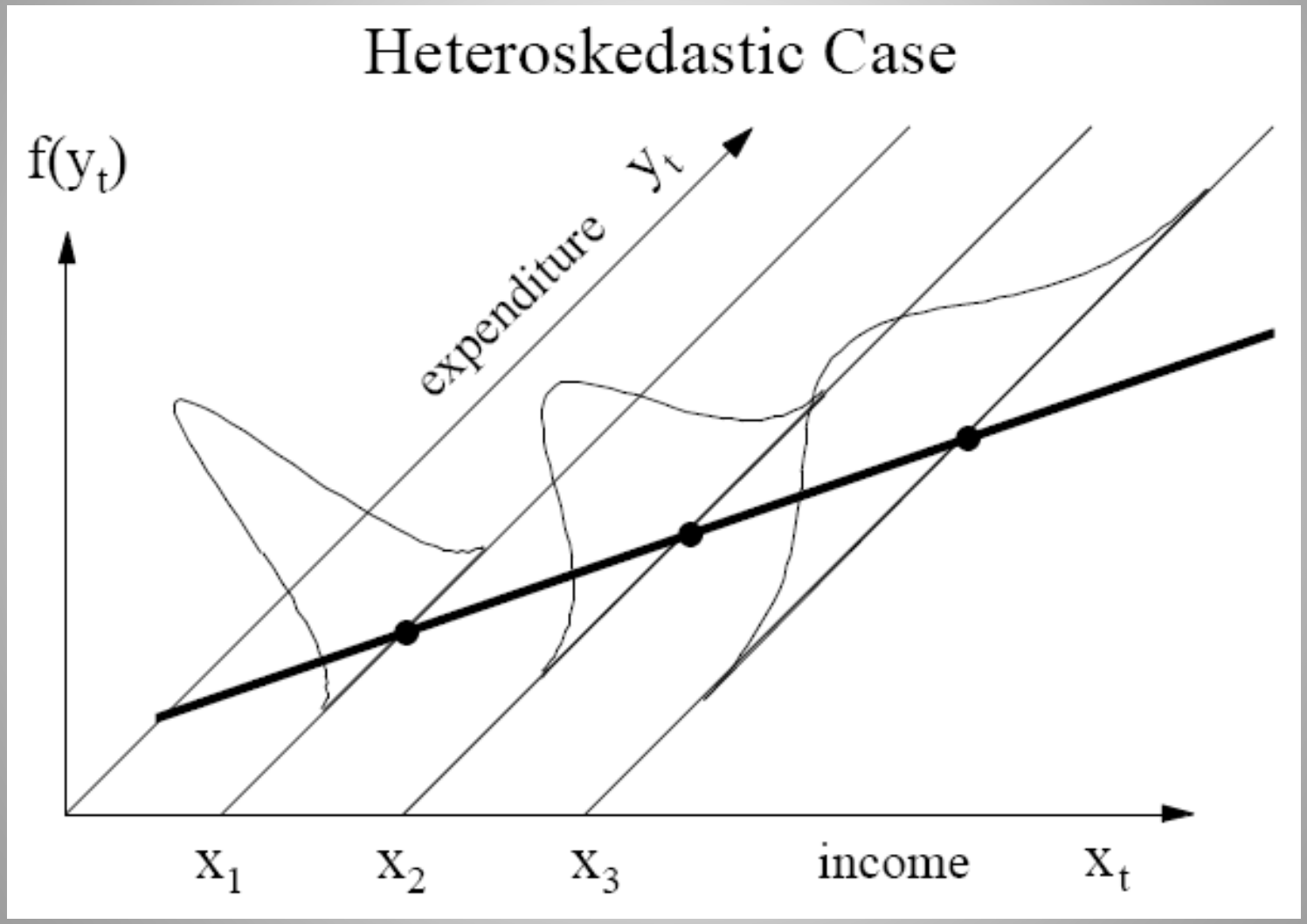
<Plot of sample data>



- 동분산 가정


Homoskedastic Case
f(y_t)
y_t
expenditure
x_1=480
x_2=800
income
x_t

- 이분산 가정



✓ 저소득 가계의 식료품 지출액과 고소득 가계의 식료품 지출액 중 어느 것을 추정하는 것이 더 쉬운가?

- 저소득 가계는 식료품에 대한 낭비적인 선택을 못함
  - 상대적으로 선택의 폭이 좁음,
    소득의 특정 부분을 식료품 지출에 사용,
    소득을 알면 식료품 지출액을 추정하기 쉬움
- 고소득 가계는 선택의 폭이 넓음 (가격보다 취향을 중시),
  소득은 설명변수로서 덜 중요하게 되고,
  소득을 알더라도 식료품 지출액을 추정하기 어려움

- OLS 추정 $y_i = \beta_1 + \beta_2 x_i + e_i$

  가정 : $E(e_i) = 0 \quad \text{var}(e_i) = \sigma^2 \quad \text{cov}(e_i, e_j) = 0$

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

- 식료품 지출액의 퍼진 정도가 소득 수준에 따라 다르다면, 동분산 가정 $\text{var}(y_i) = \text{var}(e_i) = \sigma^2$ 은 부적절함
- 다음과 같은 이분산 가정이 타당할 것임

$$\text{var}(y_i) = \text{var}(e_i) = \sigma_i^2$$

- 횡단면자료의 경우 이분산이 존재하는 경우가 흔히 있음
- 시계열자료에서도 가끔 나타남 (예: 외환위기 이후 변동성 증가)

8.2 이분산이 OLS 추정량에 미치는 영향

- OLS 추정량은 선형 불편 추정량이기는 하지만
  최소분산을 가지지는 않음 (BLUE 아님)

- 컴퓨터 SW가 계산해 주는 추정치의 표준오차는 잘못된 것

  ⇒ 이것에 근거한 가설검정 및 신뢰구간도 올바르지 않게 됨

- **OLS 추정량의 표준오차 계산식**

<동분산 가정> $y_i = \beta_1 + \beta_2 x_i + e_i$ $\quad \text{var}(e_i) = \sigma^2$

$$\text{var}(b_2) = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

<이분산 가정> $y_i = \beta_1 + \beta_2 x_i + e_i$ $\quad \text{var}(e_i) = \sigma_i^2$

$$\text{var}(b_2) = \sum_{i=1}^{N} w_i^2 \sigma_i^2 = \frac{\sum_{i=1}^{N}\left[(x_i - \bar{x})^2 \sigma_i^2\right]}{\left[\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^2}$$

- **표준오차를 정확하게 계산하기 위한 White 표준오차**

- H. White는 OLS 잔차를 이용하여,
  모수 추정량에 대한 정확한 표준오차를 계산하는 방법을 제시

$$\text{var}(b_2) = \sum_{i=1}^{N} w_i^2 \sigma_i^2 = \frac{\sum_{i=1}^{N}\left[(x_i - \bar{x})^2 \sigma_i^2\right]}{\left[\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^2}$$

$$\widehat{\text{var}(b_2)} = \sum_{i=1}^{N} w_i^2 \hat{e}_i^2 = \frac{\sum_{i=1}^{N}\left[(x_i - \bar{x})^2 \hat{e}_i^2\right]}{\left[\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^2}$$

- **White 표준오차 계산결과**

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

(27.46) (1.81) (White se)

(43.41) (2.09) (incorrect se) (OLS se)

- β_2 의 95% 신뢰구간

White: $b_2 \pm t_c \text{se}(b_2) = 10.21 \pm 2.024 \times 1.81 = [6.55,\ 13.87]$

Incorrect: (OLS) $b_2 \pm t_c \text{se}(b_2) = 10.21 \pm 2.024 \times 2.09 = [5.97,\ 14.45]$

- White 표준오차를 사용하면 신뢰구간의 폭이 더 좁게 계산됨
  (더 유용한 구간추정 결과를 알려 줌)

- OLS (동분산 가정)

Dependent Variable: FOOD_EXP
Method: Least Squares
Date: 05/11/11 Time: 14:47
Sample: 1 40
Included observations: 40

Variable	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		

- OLS (이분산 가정, White 표준오차)

Dependent Variable: FOOD_EXP
Method: Least Squares
Date: 05/11/11 Time: 14:48
Sample: 1 40
Included observations: 40
White Heteroskedasticity-Consistent Standard Errors & Covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	83.41600	27.46375	3.037313	0.0043
INCOME	10.20964	1.809077	5.643565	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		

8.3 일반 최소제곱 추정량 $y_i = \beta_1 + \beta_2 x_i + e_i, \quad \mathrm{var}(e_i) = \sigma_i^2$

- 이분산이 존재하는 경우에 적용할 수 있는 최선의 추정방법
  $\Rightarrow$ GLS (Generalized Least Squares) (BLUE 임)

- 가정을 $\mathrm{var}(y_i) = \mathrm{var}(e_i) = \sigma_i^2$ 으로 수정하는 것만으로는 회귀식의 모수를 추정할 수 없음
  - N개의 표본만으로 N개의 상이한 분산 $(\sigma_1^2, \sigma_2^2, \cdots, \sigma_N^2)$ 과 모수들 (β_1, β_2) 을 추정하는 것은 불가능 (자유도 부족)

- 이 문제를 극복하기 위해서는
  σ_i^2 에 대한 추가적인 가정이 필요

Two Types of Heteroskedasticity

1. **Proportional** Heteroskedasticity.
   (continuous function(of x_t, for example))

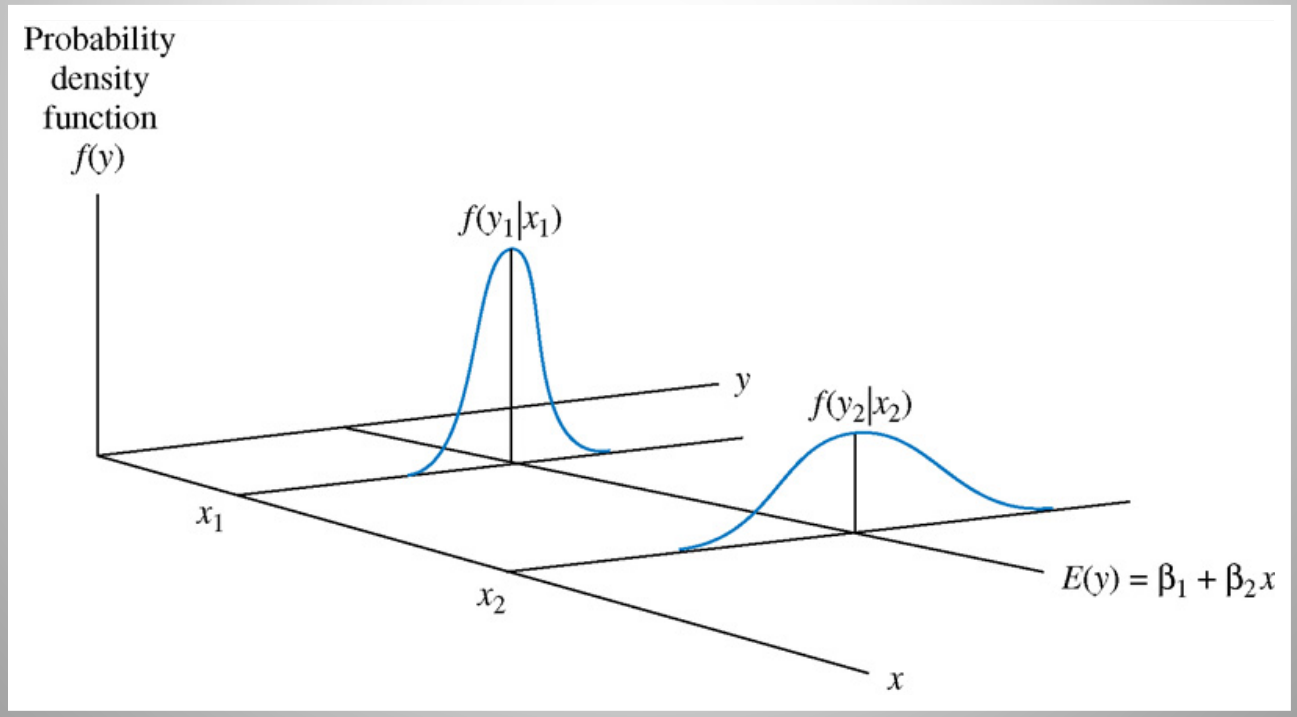
2. **Partitioned** Heteroskedasticity.
   (discrete categories/groups)

8.3.1 비례적 이분산

- 식료품 지출액의 경우 오차 분산은 소득 수준에 비례

  ⇒ 비례적 이분산을 가정하는 것이 적절함, 다음은 한 사례

$$\mathrm{var}(e_i) = \sigma_i^2 = \sigma^2 x_i$$



- 이분산이 존재하는 경우 OLS 추정량은 BLUE가 아님

⇒ 이 문제를 해결하는 한 방법은 회귀모형을 변형시켜 동분산 오차를 갖도록 한 다음, OLS를 적용하는 것

- 회귀모형 $y_i = \beta_1 + \beta_2 x_i + e_i$ 의 양변을 $\sqrt{x_i}$ 로 나누어 보자

$$\frac{y_i}{\sqrt{x_i}} = \beta_1 \frac{1}{\sqrt{x_i}} + \beta_2 \frac{x_i}{\sqrt{x_i}} + \frac{e_i}{\sqrt{x_i}}$$

- 다음과 같이 변형된 변수를 정의하면,

$$y_i^* = \frac{y_i}{\sqrt{x_i}} \quad x_{i1}^* = \frac{1}{\sqrt{x_i}} \quad x_{i2}^* = \frac{x_i}{\sqrt{x_i}} \quad e_i^* = \frac{e_i}{\sqrt{x_i}}$$

- 변형된 모형은 다음과 같음 (상수항이 없는 식이라는 점에 유의)

$$y_i^* = \beta_1 x_{i1}^* + \beta_2 x_{i2}^* + e_i^*$$

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

$$\frac{y_t}{\sqrt{x_t}} = \beta_1 \frac{1}{\sqrt{x_t}} + \beta_2 \frac{x_t}{\sqrt{x_t}} + \frac{e_t}{\sqrt{x_t}}$$

$$y_t^* = \beta_1 x_{t1}^* + \beta_2 x_{t2}^* + e_t^*$$

e_t is **heteroskedastic**, but e_t^* is **homoskedastic**

- 변형된 모형의 장점은 오차항 e_i^* 가 동분산을 가진다는 것

$$y_i^* = \beta_1 x_{i1}^* + \beta_2 x_{i2}^* + e_i^* \qquad e_i^* = \frac{e_i}{\sqrt{x_i}} \qquad \text{var}(e_i) = \sigma_i^2 = \sigma^2 x_i$$

(증명) $$\text{var}(e_i^*) = \text{var}\left(\frac{e_i}{\sqrt{x_i}}\right) = \frac{1}{x_i}\text{var}(e_i) = \frac{1}{x_i}\sigma^2 x_i = \sigma^2$$

- 오차항의 평균도 0, 독립성 가정도 충족됨

$$E(e_i^*) = 0 \qquad \text{cov}(e_i^*, e_j^*) = 0$$

- 변형된 변수 $(y_i^*,\ x_{i1}^*,\ x_{i2}^*)$ 를 이용하여 OLS로 추정하면, 모수 (β_1, β_2) 에 대한 BLUE를 구할 수 있음
- 이렇게 구한 추정량을 generalized least squares (GLS) estimator 혹은 weighted least squares (WLS) estimator 라고 함

▪ SAS program

```
data food2 ;                              * create dataset;
infile 'C:₩tmp₩table3-1.prn' ;            * read in data=food;
input y x ;                               * input variables;

w = 1/x ;                                 * create weight variable;


proc reg ;                          * estimate regression;
food_GLS2 : model y=x ;             * use original data;
weight w ;                          * specify weight for weighted LS = GLS;
run ;
```

<식료품 지출액 사례>

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

  (27.46) (1.81) (White se)

  (43.41) (2.09) (incorrect se) (OLS se)

- WLS 추정결과 $\hat{y}_i = 78.68 + 10.45 x_i$

  (se) (23.79) (1.39)

- WLS 추정치의 95% 신뢰구간

$$\hat{\beta}_2 \pm t_c \mathrm{se}(\hat{\beta}_2) = 10.451 \pm 2.024 \times 1.386 = [7.65, 13.26]$$

- WLS

Dependent Variable: FOOD_EXP
Method: Least Squares
Date: 05/12/11 Time: 14:57
Sample: 1 40
Included observations: 40
Weighting series: 1/INCOME^(0.5)

$$\hat{y}_i = 78.68 + 10.45x_i$$
$$(se) \quad (23.79) \quad (1.39)$$

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	78.68408	23.78872	3.307621	0.0021
INCOME	10.45101	1.385891	7.541002	0.0000

Weighted Statistics

R-squared	0.599438	Mean dependent var	263.3689
Adjusted R-squared	0.588897	S.D. dependent var	76.43899
S.E. of regression	76.30741	Akaike info criterion	11.55612
Sum squared resid	221267.2	Schwarz criterion	11.64057
Log likelihood	-229.1225	Hannan-Quinn criter.	11.58666
F-statistic	56.86672	Durbin-Watson stat	1.905701
Prob(F-statistic)	0.000000		

Unweighted Statistics

R-squared	0.384787	Mean dependent var	283.5735
Adjusted R-squared	0.368597	S.D. dependent var	112.6752
S.E. of regression	89.53266	Sum squared resid	304611.7
Durbin-Watson stat	1.892377		

Quick-Estimate Equation-Options-WLS 클릭-Weight 부분에 1/income^(0.5) 입력

$$\hat{y}_i = 83.42 + 10.21x_i$$
$$(43.41) \quad (2.09)$$

8.3.2 분산함수의 추정

- 식료품 지출액 모형에서 이분산을 나타낼 수 있는 방법들

$$\mathrm{var}(e_i) = \sigma_i^2 = \sigma^2 x_i \qquad \mathrm{var}(e_i) = \sigma_i^2 = \sigma^2 x_i^2 \qquad \mathrm{var}(e_i) = \sigma_i^2 = \sigma^2 x_i^{1/2}$$

- 위의 가정들을 일반화하면 아래와 같이 나타낼 수 있음

$$\mathrm{var}(e_i) = \sigma_i^2 = \sigma^2 x_i^{\gamma}$$ ($\gamma = 1$ 로 가정할 필요 없음)

- 이식을 약간 변형하면, 위 가정을 아래와 같이 쓸 수 있음

$$\ln(\sigma_i^2) = \ln(\sigma^2) + \gamma \ln(x_i)$$

$$\sigma_i^2 = \exp\left(\ln(\sigma^2) + \gamma \ln(x_i)\right)$$

$$= \exp(\alpha_1 + \alpha_2 z_i)$$

$$\alpha_1 = \ln(\sigma^2)$$

$$\alpha_2 = \gamma$$

$$z_i = \ln(x_i)$$

- **일반적인 분산함수**

- $\sigma_i^2 = \exp(\alpha_1 + \alpha_2 z_i)$ 보다 더 일반화된 분산함수

$$\sigma_i^2 = \exp(\alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_s z_{iS})$$

여기서 z_{ik}는 분산과 관련된 설명변수들

- 다음과 같이 표현해도 됨

$$\ln(\sigma_i^2) = \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_s z_{iS}$$

- **분산함수의 추정 방법**

- 아래와 같은 간단한 경우로 설명해 보자.

$$y_i = E(y_i) + e_i = \beta_1 + \beta_2 x_i + e_i \qquad \ln(\sigma_i^2) = \alpha_1 + \alpha_2 z_i$$

(1) OLS 추정하여 잔차 $\hat{e}_i$ 를 구함 $\hat{e}_i = y_i - \hat{y}_i = y_i - b_1 - b_2 x_i$

(2) z_i 를 선택하여 아래 분산함수를 추정

$$\ln(\hat{e}_i^2) = \ln(\sigma_i^2) + v_i = \alpha_1 + \alpha_2 z_i + v_i$$

식료품 지출액 사례: $\ln(\hat{\sigma}_i^2) = 0.9378 + 2.329 \ln(x_i)$

$$\gamma = 2.329 \neq 1$$

(3) 분산의 추정

$$\hat{\sigma}_i^2 = \exp(\hat{\alpha}_1 + \hat{\alpha}_2 z_i)$$

- 분산함수의 추정

Dependent Variable: LOG(EHAT2)
Method: Least Squares
Date: 05/13/11 Time: 13:54
Sample: 1 40
Included observations: 40
LOG(EHAT2)=C(1)+C(2)*LOG(INCOME)

$$\ln(\hat{\sigma}_i^2) = 0.9378 + 2.329\ln(x_i)$$

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	0.937796	1.583106	0.592377	0.5571
C(2)	2.329239	0.541336	4.302761	0.0001

R-squared	0.327597	Mean dependent var	7.648159
Adjusted R-squared	0.309903	S.D. dependent var	2.071519
S.E. of regression	1.720855	Akaike info criterion	3.972226
Sum squared resid	112.5310	Schwarz criterion	4.056670
Log likelihood	-77.44452	Hannan-Quinn criter.	4.002758
F-statistic	18.51375	Durbin-Watson stat	2.175575
Prob(F-statistic)	0.000114		

- **모수의 GLS 추정 방법** $y_i = \beta_1 + \beta_2 x_i + e_i$

- 앞에서 구한 분산 σ_i 를 이용하여 GLS 추정량 구하면 됨

(1) 회귀모형의 양변을 σ_i 로 나눔

$$\left(\frac{y_i}{\sigma_i}\right) = \beta_1\left(\frac{1}{\sigma_i}\right) + \beta_2\left(\frac{x_i}{\sigma_i}\right) + \left(\frac{e_i}{\sigma_i}\right)$$

- 변형된 오차는 동분산임을 확인할 수 있음

$$\text{var}\left(\frac{e_i}{\sigma_i}\right) = \left(\frac{1}{\sigma_i^2}\right)\text{var}(e_i) = \left(\frac{1}{\sigma_i^2}\right)\sigma_i^2 = 1$$

(2) $\hat{\sigma}_i$을 이용하여 변형된 변수를 계산함

$$y_i^* = \left(\frac{y_i}{\hat{\sigma}_i}\right), \quad x_{i1}^* = \left(\frac{1}{\hat{\sigma}_i}\right), \quad x_{i2}^* = \left(\frac{x_i}{\hat{\sigma}_i}\right)$$

(3) 아래 식을 OLS로 추정함 (BLUE)

$$y_i^* = \beta_1 x_{i1}^* + \beta_2 x_{i2}^* + e_i^*$$

- **GLS 추정절차 요약** $y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_K x_{iK} + e_i$

(1) OLS로 추정하여 잔차 $\hat{e}_i^2$ 을 계산함

(2) $\ln \hat{e}_i^2 = \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS} + v_i$ 에 OLS 적용하여 $\alpha_1, \alpha_2, \ldots, \alpha_S$ 추정

(3) 분산 추정값 계산 $\hat{\sigma}_i^2 = \exp(\hat{\alpha}_1 + \hat{\alpha}_2 z_{i2} + \cdots + \hat{\alpha}_S z_{iS})$

(4) 회귀모형의 양변을 $\hat{\sigma}_i$ 로 나누어, 변형된 자료 계산

$$y_i^* = \left(\frac{y_i}{\hat{\sigma}_i}\right), \quad x_{i1}^* = \left(\frac{1}{\hat{\sigma}_i}\right), \quad x_{ik}^* = \left(\frac{x_{ik}}{\hat{\sigma}_i}\right)$$

(5) 변형된 아래 모형을 OLS로 추정

$$y_i^* = \beta_1 x_{i1}^* + \beta_2 x_{i2}^* + \beta_3 x_{i3}^* + \cdots + \beta_K x_{iK}^* + e_i^*$$

<식료품 지출액 사례> $\text{var}(e_i) = \sigma_i^2 = \sigma^2 x_i^\gamma$

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

  $\gamma = 0$

  (27.46) (1.81) (White se)

  (43.41) (2.09) (OLS se)

- WLS 추정결과 $\hat{y}_i = 78.68 + 10.45 x_i$

  $\gamma = 1$

  (se) (23.79) (1.39)

- GLS 추정결과 $\hat{y}_i = 76.05 + 10.63 x$

  $\gamma = 2.329$

  (se) (9.71) (0.97)

- GLS

Dependent Variable: FOOD_EXP
Method: Least Squares
Date: 05/12/11 Time: 17:14
Sample: 1 40
Included observations: 40
Weighting series: 1/(INCOME^2.329)^(0.5)

$$\hat{y}_i = 76.05 + 10.63x$$

(se) (9.71) (0.97)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	76.05387	9.714921	7.828563	0.0000
INCOME	10.63348	0.971543	10.94494	0.0000

Weighted Statistics

R-squared	0.759176	Mean dependent var	225.3697
Adjusted R-squared	0.752839	S.D. dependent var	105.2968
S.E. of regression	55.59375	Akaike info criterion	10.92273
Sum squared resid	117445.3	Schwarz criterion	11.00717
Log likelihood	-216.4545	Hannan-Quinn criter.	10.95326
F-statistic	119.7918	Durbin-Watson stat	1.905316
Prob(F-statistic)	0.000000		

Unweighted Statistics

R-squared	0.384266	Mean dependent var	283.5735
Adjusted R-squared	0.368063	S.D. dependent var	112.6752
S.E. of regression	89.57055	Sum squared resid	304869.6
Durbin-Watson stat	1.890159		

8.3.3 이분산적 분할(Heteroskedastic Partition)

(예) 임금함수

임금 = f(교육수준, 경험, 인종,성별, 거주지역, ...)

- 단순화를 위해 임금($WAGE$)이 다음 세 가지에만 의존한다고 하자
  - 교육받은 연수(EDUC)
  - 경험의 연수(숙련도, 생산성의대리변수)($EXPER$)
  - 거주지역($METRO$, 더미변수: 대도시 거주면 1, 아니면 0)

- OLS 추정결과

$$\widehat{WAGE} = -9.914 + 1.234EDUC + 0.133EXPER + 1.524METRO$$
$$(se) \qquad (1.08) \quad (0.070) \qquad (0.015) \qquad (0.431)$$

  - 대도시 지역 평균임금이 시골지역보다 시간당 $1.524 더 높음

- 임금함수 (OLS)

$$\widehat{WAGE} = -9.914 + 1.234EDUC + 0.133EXPER + 1.524METRO$$
$$(se) \qquad (1.08) \quad (0.070) \qquad (0.015) \qquad (0.431)$$

Dependent Variable: WAGE
Method: Least Squares
Date: 05/12/11 Time: 17:32
Sample: 1 1000
Included observations: 1000
WAGE=C(1)+C(2)*EDUC+C(3)*EXPER+C(4)*METRO

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	-9.913984	1.075663	-9.216631	0.0000
C(2)	1.233964	0.069961	17.63782	0.0000
C(3)	0.133244	0.015232	8.747835	0.0000
C(4)	1.524104	0.431091	3.535459	0.0004

R-squared	0.266903	Mean dependent var	10.21302
Adjusted R-squared	0.264695	S.D. dependent var	6.246641
S.E. of regression	5.356490	Akaike info criterion	6.198487
Sum squared resid	28577.21	Schwarz criterion	6.218118
Log likelihood	-3095.243	Hannan-Quinn criter.	6.205948
F-statistic	120.8733	Durbin-Watson stat	0.502560
Prob(F-statistic)	0.000000		

❖ 대도시지역(M)과 시골지역(R)의 임금 분산은 다를 수 있음

- 대도시에는 다양한 형태의 직업이 존재, 임금 분산이 더 클 가능성

- 두 지역에서 교육수준과 경험이 임금에 미치는 영향은 동일하지만 임금 분산은 다르다고 가정해보자

$$WAGE_{Mi} = \beta_{M1} + \beta_2 EDUC_{Mi} + \beta_3 EXPER_{Mi} + e_{Mi} \qquad i = 1, 2, \ldots, N_M \ (808)$$

$$WAGE_{Ri} = \beta_{R1} + \beta_2 EDUC_{Ri} + \beta_3 EXPER_{Ri} + e_{Ri} \qquad i = 1, 2, \ldots, N_R \ (192)$$

- 두 식의 오차 분산이 동일하다면 $(\text{var}(e_{Mi}) = \text{var}(e_{Ri}) = \sigma^2)$

OLS로 각각 분리 추정하였을 때 두 식의 (β_2, β_3)은 동일하고 두 식의 상수항은 다음의 관계에 있을 것임 (실제로는 다름. Why?)

$$b_{M1} = b_{R1} + 1.524 = -9.914 + 1.524 = -8.39$$

- 대도시지역과 시골지역의 임금 분산이 다르다고 가정하면
  (즉, 이분산적 분할이 존재한다면), 다음과 같이 나타낼 수 있음

$$\mathrm{var}(e_{Mi}) = \sigma_M^2, \qquad \mathrm{var}(e_{Ri}) = \sigma_R^2$$

- 대도시 808개, 시골 192개 표본을 이용하여
  앞의 두 식을 각각 분리 추정한 결과는 다음과 같음

$$\hat{\sigma}_M^2 = 31.824, \qquad \hat{\sigma}_R^2 = 15.243$$

$b_{M1} = -9.052$	$b_{M2} = 1.282$	$b_{M3} = 0.1346$
$b_{R1} = -6.166$	$b_{R2} = 0.956$	$b_{R3} = 0.1260$

- 교육과 경험이 임금에 미치는 영향이 어느 정도인지 알 수 없음

✓ 전체 표본을 이용하여 GLS로 추정하는 것이 바람직함

- 임금함수 (도시 808명, metro=1)

Dependent Variable: WAGE
Method: Least Squares
Date: 05/13/11 Time: 17:25
Sample: 193 1000
Included observations: 808

$$\text{var}(e_{Mi}) = \sigma_M^2 = \frac{\sum \hat{e}^2}{N-K} = \frac{25618.1}{(808-3)} = 31.824$$

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-9.052478	1.189456	-7.610603	0.0000
EDUC	1.281714	0.079763	16.06910	0.0000
EXPER	0.134560	0.017948	7.497370	0.0000

R-squared	0.258183	Mean dependent var	10.57802
Adjusted R-squared	0.256340	S.D. dependent var	6.541667
S.E. of regression	5.641253	Akaike info criterion	6.301795
Sum squared resid	25618.10	Schwarz criterion	6.319226
Log likelihood	-2542.925	Hannan-Quinn criter.	6.308488
F-statistic	140.0868	Durbin-Watson stat	0.477626
Prob(F-statistic)	0.000000		

- 임금함수 (농촌 192명, metro=0)

Dependent Variable: WAGE
Method: Least Squares
Date: 05/13/11 Time: 17:24
Sample: 1 192
Included observations: 192

$$\text{var}(e_{Ri}) = \sigma_R^2 = \frac{\sum \hat{e}^2}{N-K} = \frac{2880.924}{(192-3)} = 15.243$$

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-6.165855	1.898511	-3.247732	0.0014
EDUC	0.955585	0.133190	7.174608	0.0000
EXPER	0.125974	0.024771	5.085538	0.0000

R-squared	0.258748	Mean dependent var	8.676979
Adjusted R-squared	0.250904	S.D. dependent var	4.510933
S.E. of regression	3.904227	Akaike info criterion	5.577498
Sum squared resid	2880.924	Schwarz criterion	5.628397
Log likelihood	-532.4398	Hannan-Quinn criter.	5.598112
F-statistic	32.98706	Durbin-Watson stat	0.516503
Prob(F-statistic)	0.000000		

- **이분산적 분할이 존재하는 모형의 추정방법**

(1) 두 하위표본(대도시/시골)에 대해 각각 추정하여 분산 계산

(2) 전체 표본을 각 그룹의 오차항 표준오차로 나누어 변수 변형

$$\left(\frac{WAGE_i}{\hat{\sigma}_i}\right)=\beta_{R1}\left(\frac{1}{\hat{\sigma}_i}\right)+\beta_2\left(\frac{EDUC_i}{\hat{\sigma}_i}\right)+\beta_3\left(\frac{EXPER_i}{\hat{\sigma}_i}\right)+\delta\left(\frac{METRO_i}{\hat{\sigma}_i}\right)+\left(\frac{e_i}{\hat{\sigma}_i}\right)$$

$$\hat{\sigma}_i=\begin{cases}\hat{\sigma}_M & \text{when } METRO_i=1\\ \hat{\sigma}_R & \text{when } METRO_i=0\end{cases}$$

(3) 변형된 모형에 OLS 적용 (BLUE)

- **임금함수 추정결과 비교**

<동분산 가정하여 OLS 적용한 결과>

$$\widehat{WAGE} = -9.914 + 1.234EDUC + 0.133EXPER + 1.524METRO$$

(se) (1.08) (0.070) (0.015) (0.431)

<이분산 가정하여 GLS 적용한 결과>

$$\widehat{WAGE} = -9.398 + 1.196EDUC + 0.132EXPER + 1.539METRO$$

(se) (1.02) (0.069) (0.015) (0.346)

- 모수 추정치는 비슷함, 이는 이분산이 존재하는 경우에도 OLS 추정량은 불편추정량이기 때문임 (GLS 추정치가 더 정확함)
- GLS 경우의 표준오차가 별로 줄지 않은 이유는 표본이 작기 때문임

- 임금함수 (OLS)

$$\widehat{WAGE} = -9.914 + 1.234EDUC + 0.133EXPER + 1.524METRO$$

$$(se) \quad (1.08) \quad (0.070) \quad (0.015) \quad (0.431)$$

Dependent Variable: WAGE
Method: Least Squares
Date: 05/12/11 Time: 17:32
Sample: 1 1000
Included observations: 1000
WAGE=C(1)+C(2)*EDUC+C(3)*EXPER+C(4)*METRO

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	-9.913984	1.075663	-9.216631	0.0000
C(2)	1.233964	0.069961	17.63782	0.0000
C(3)	0.133244	0.015232	8.747835	0.0000
C(4)	1.524104	0.431091	3.535459	0.0004

R-squared	0.266903	Mean dependent var	10.21302
Adjusted R-squared	0.264695	S.D. dependent var	6.246641
S.E. of regression	5.356490	Akaike info criterion	6.198487
Sum squared resid	28577.21	Schwarz criterion	6.218118
Log likelihood	-3095.243	Hannan-Quinn criter.	6.205948
F-statistic	120.8733	Durbin-Watson stat	0.502560
Prob(F-statistic)	0.000000		

- 임금함수 (GLS)

$$\widehat{WAGE} = -9.398 + 1.196EDUC + 0.132EXPER + 1.539METRO$$
$$(se) \qquad (1.02) \quad (0.069) \qquad (0.015) \qquad (0.346)$$

Dependent Variable: WA_S
Method: Least Squares
Date: 05/14/11 Time: 16:43
Sample: 1 1000
Included observations: 1000
WA_S=C(1)*CONS_S+C(2)*ED_S+C(3)*EX_S+C(4)*ME_S

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	-9.398355	1.019672	-9.217037	0.0000
C(2)	1.195720	0.068508	17.45375	0.0000
C(3)	0.132209	0.014548	9.087453	0.0000
C(4)	1.538803	0.346285	4.443749	0.0000

R-squared	0.265251	Mean dependent var	1.941801
Adjusted R-squared	0.263038	S.D. dependent var	1.166283
S.E. of regression	1.001213	Akaike info criterion	2.844294
Sum squared resid	998.4178	Schwarz criterion	2.863925
Log likelihood	-1418.147	Hannan-Quinn criter.	2.851755
Durbin-Watson stat	0.510038		

8.4 이분산의 탐지

- GLS를 적용해야 할 정도로 이분산이 문제되는지를 탐지하는 방법

8.4.1 잔차의 도표화

- 단순회귀 경우 ⇒ plot을 그려보는 방법 (예) 식료품 지출액
- 다중회귀 경우 ⇒ 각 설명변수와 잔차의 plot을 그려보는 방법

✓ 그래프만으로 판단하기 곤란할 수도 있음

- 일반적인 방법 ⇒ 골드펠드-콴트 검정

(Goldfeld-Quandt test, GQ test)

8.4.2 Goldfeld-Quandt test

- 전체 표본을 상이한 분산을 갖고 있을 것으로 예상되는 2개의 그룹(하위표본)으로 나눔

  - 비례적 이분산 경우 $\Rightarrow$ 잔차의 크기 순으로 두 그룹으로 구분 (다음 슬라이드에서 설명)

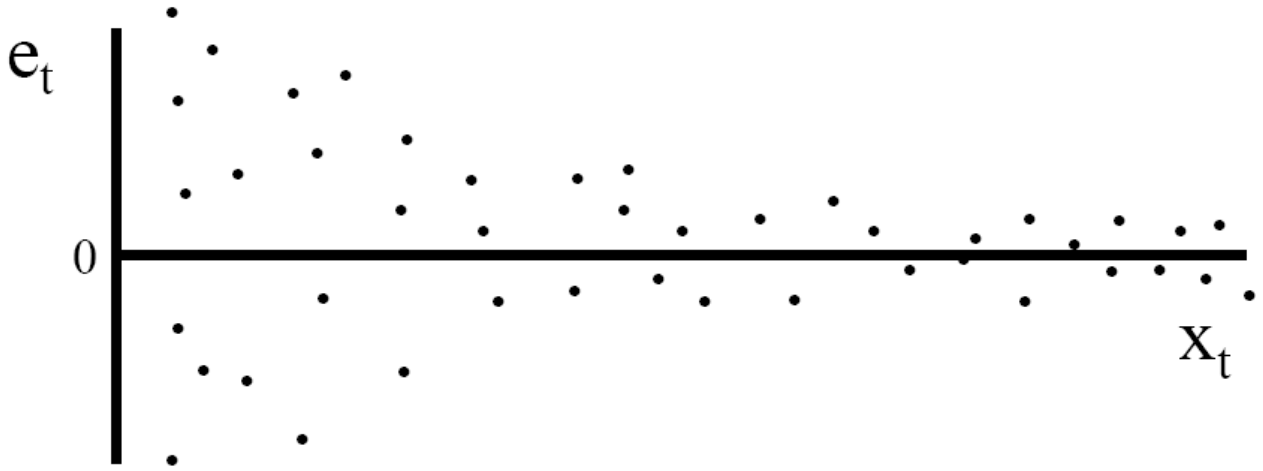
  - 이분산적 분할의 경우 $\Rightarrow$ 쉽게 구분 가능 (대도시/시골)

- 다음의 가설을 검정

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ against } H_1 : \sigma_1^2 \neq \sigma_2^2$$

Residual Plots

Plot residuals against one variable at a time after sorting the data by that variable to try to find a heteroskedastic pattern in the data.



- 임금함수 (도시 808명, metro=1)

Dependent Variable: WAGE
Method: Least Squares
Date: 05/13/11 Time: 17:25
Sample: 193 1000
Included observations: 808

$$\text{var}(e_{Mi}) = \sigma_M^2 = \frac{\sum \hat{e}^2}{N-K} = \frac{25618.1}{(808-3)} = 31.824$$

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-9.052478	1.189456	-7.610603	0.0000
EDUC	1.281714	0.079763	16.06910	0.0000
EXPER	0.134560	0.017948	7.497370	0.0000

R-squared	0.258183	Mean dependent var	10.57802
Adjusted R-squared	0.256340	S.D. dependent var	6.541667
S.E. of regression	5.641253	Akaike info criterion	6.301795
Sum squared resid	25618.10	Schwarz criterion	6.319226
Log likelihood	-2542.925	Hannan-Quinn criter.	6.308488
F-statistic	140.0868	Durbin-Watson stat	0.477626
Prob(F-statistic)	0.000000		

- 임금함수 (농촌 192명, metro=0)

Dependent Variable: WAGE
Method: Least Squares
Date: 05/13/11 Time: 17:24
Sample: 1 192
Included observations: 192

$$\text{var}(e_{Ri}) = \sigma_R^2 = \frac{\sum \hat{e}^2}{N-K} = \frac{2880.924}{(192-3)} = 15.243$$

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-6.165855	1.898511	-3.247732	0.0014
EDUC	0.955585	0.133190	7.174608	0.0000
EXPER	0.125974	0.024771	5.085538	0.0000

R-squared	0.258748	Mean dependent var	8.676979
Adjusted R-squared	0.250904	S.D. dependent var	4.510933
S.E. of regression	3.904227	Akaike info criterion	5.577498
Sum squared resid	2880.924	Schwarz criterion	5.628397
Log likelihood	-532.4398	Hannan-Quinn criter.	5.598112
F-statistic	32.98706	Durbin-Watson stat	0.516503
Prob(F-statistic)	0.000000		

- **임금함수 경우의 Goldfeld-Quandt Test**

- 가설: $H_0 : \sigma_M^2 = \sigma_R^2$ against $H_1 : \sigma_M^2 > \sigma_R^2$

- 검정통계량: $$F = \frac{\hat{\sigma}_M^2}{\hat{\sigma}_R^2} \sim F_{(N_M - K_M, N_R - K_R)}$$

- 임계치: 유의수준 5% 경우 $F_C = F_{(\alpha; N_M - K_M, N_R - K_R)} = F_{(0.95; 808-3, 192-3)} = 1.22$

- 검정결과: $$F = \frac{\hat{\sigma}_M^2}{\hat{\sigma}_R^2} = \frac{31.824}{15.243} = 2.09 > F_C = 1.22$$

- 귀무가설 기각, 대도시의 임금 분산이 시골보다 더 큼(이분산)

- **식료품 지출액 경우의 Goldfeld-Quandt Test**

(1) 표본을 비슷한 크기의 두 그룹으로 분리함

(2) 큰 분산을 가질 것 같은 그룹(1)의 분산 추정값 $\Rightarrow \hat{\sigma}_1^2$

작은 분산을 가질 것 같은 그룹(2)의 분산 추정값 $\Rightarrow \hat{\sigma}_2^2$

(3) 가설 : $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 > \sigma_2^2$

(4) 검정통계량: $GQ = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \sim F_{(N_1-K),\,(N_2-K)}$ (이분산이 심할수록 GQ는 큰 값)

(5) 자유도 (18,18)인 F-분포에서 5% 임계값은 $F_c = 2.22$

(6) $\hat{\sigma}_1^2 = 12921.9 \quad \hat{\sigma}_2^2 = 3574.8 \quad GQ = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{12921.9}{3574.8} = 3.61$

- 귀무가설 기각 $\Rightarrow$ 이분산 존재, 오차항 분산은 소득수준에 의존함

8.4.3 분산함수의 검정

- 앞에서 분산함수를 다음과 같이 나타내었음

$$\text{var}(e_i) = \sigma_i^2 = \exp(\alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_s z_{iS})$$

- 이보다 더 일반적인 경우는 다음과 같음

$$\text{var}(y_i) = \text{var}(e_i) = E(e_i^2) = \sigma_i^2 = h(\alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS})$$

- 앞에서 사용한 분산함수들은 이것의 특정한 형태로 볼 수 있음

$$h(\alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS}) = \exp(\alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS})$$

$$h(\alpha_1 + \alpha_2 z_i) = \exp\left(\ln(\sigma^2) + \gamma \ln(x_i)\right)$$

- 일반적인 분산함수

$$\text{var}(y_i) = \sigma_i^2 = E(e_i^2) = h(\alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS})$$

- 이 일반적인 분산함수의 다른 특별한 예는 선형함수 형태임

$$\sigma_i^2 = E(e_i^2) = \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_s z_{iS}$$

- 이분산 여부를 알아보기 위해서는 다음의 가설검정을 수행하면 됨

$$H_0 : \alpha_2 = \alpha_3 = \cdots = \alpha_S = 0$$

$$H_1 : \text{ not all the } \alpha_s \text{ in } H_0 \text{ are zero}$$

- 만약 귀무가설이 옳다면, 오차 분산은 다음과 같은 동분산 형태가 됨

$$\sigma_i^2 = E(e_i^2) = \alpha_1$$

- 귀무가설이 기각되면, 오차 분산은 이분산 형태가 될 것임

- **분산함수 형태에 대한 가설검정 (이분산 검정)**

  = Lagrange multiplier test 혹은 Breusch-Pagan test

(1) 다음의 회귀식을 OLS로 추정

$$\hat{e}_i^2 = \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS} + v_i$$

(2) 가설 $H_0 : \alpha_2 = \alpha_3 = \cdots = \alpha_S = 0$

H_1 : not all the α_s in H_0 are zero

(3) 아래의 검정통계량을 이용해 검정

$$\chi^2 = N \times R^2 \sim \chi^2_{(S-1)}$$

▪ 이분산에 관한 White test

- 앞의 Lagrange multiplier test (혹은 Breusch-Pagan test)를 사용하기 쉽도록 단순화시킨 검정 방법
- z를 설명변수 x, x의 제곱항, x들의 교차곱항(상호작용항)으로 정의함

$$E(y_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}$$

$$z_2 = x_2, \quad z_3 = x_3, \quad z_4 = x_2^2, \quad z_5 = x_3^2, \quad (z_6 = x_2 x_3)$$

(예) 식료품 지출액 모형의 이분산 검정 (1) $y_i = \beta_1 + \beta_2 x_i + e_i$

< Lagrange multiplier test >

- 분산함수: $\sigma_i^2 = E(e_i^2) = \alpha_1 + \alpha_2 z_i$
- 가설: $H_0 : \alpha_2 = 0 \quad H_1 : \alpha_2 \neq 0$
- OLS 이용 $\hat{e}_i^2 = \alpha_1 + \alpha_2 x_i + v_i$ 를 추정

$$SST = 4,610,749,441 \qquad SSE = 3,759,556,169$$

$$R^2 = 1 - \frac{SSE}{SST} = 0.1846$$

$$\chi^2 = N \times R^2 = 40 \times 0.1846 = 7.38 > \chi^2(1, 0.95) = 3.84$$

- 유의수준 5%에서 귀무가설 기각, 따라서 이분산 존재함

- 식료품 이분산 LM 검정 (1차항 이분산과 관련되는 경우)

Dependent Variable: EHAT2
Method: Least Squares
Date: 05/14/11 Time: 16:53
Sample: 1 40
Included observations: 40
EHAT2=C(1)+C(2)*INCOME

$$\hat{e}_i^2 = \alpha_1 + \alpha_2 x_i + v_i$$

$$\chi^2 = N \times R^2 = 40 \times 0.1846 = 7.38$$

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	-5762.370	4823.501	-1.194645	0.2396
C(2)	682.2326	232.5920	2.933172	0.0057

R-squared	0.184611	Mean dependent var	7612.629
Adjusted R-squared	0.163153	S.D. dependent var	10873.10
S.E. of regression	9946.642	Akaike info criterion	21.29656
Sum squared resid	3.76E+09	Schwarz criterion	21.38101
Log likelihood	-423.9313	Hannan-Quinn criter.	21.32710
F-statistic	8.603501	Durbin-Watson stat	2.344523
Prob(F-statistic)	0.005659		

- 식료품 이분산 Breusch-Pagan test 검정 / Eviews output (1차항 이분산과 관련되는 경우)

$$\chi^2 = N \times R^2 = 40 \times 0.1846 = 7.38$$

Heteroskedasticity Test: Breusch-Pagan-Godfrey

F-statistic	8.603501	Prob. F(1,38)	0.0057
Obs*R-squared	7.384424	Prob. Chi-Square(1)	0.0066
Scaled explained SS	6.627901	Prob. Chi-Square(1)	0.0100

(예) 식료품 지출액 모형의 이분산 검정 (2) $y_i = \beta_1 + \beta_2 x_i + e_i$

< White test >

- 분산함수: $\sigma_i^2 = E(e_i^2) = \alpha_1 + \alpha_2 x_i + \alpha_3 x_i^2$
- 가설: $H_0 : \alpha_2 = \alpha_3 = 0, \quad H_1 : \alpha_2 \neq 0 \text{ or } \alpha_3 \neq 0$
- OLS 이용 $\hat{e}_i^2 = \alpha_1 + \alpha_2 x_i + \alpha_3 x_i^2 + v_i$ 를 추정

$$\chi^2 = N \times R^2 = 40 \times 0.18888 = 7.555 > \chi^2(2, 0.95) = 5.99$$

$$p\text{-value} = 0.023$$

- 유의수준 5%에서 귀무가설 기각, 따라서 이분산 존재함

- 식료품 이분산 White 검정 (2차항까지 이분산과 관련되는 경우)

Heteroskedasticity Test: Breusch-Pagan-Godfrey

F-statistic	4.307884	Prob. F(2,37)	0.0208
Obs*R-squared	7.555079	Prob. Chi-Square(2)	0.0229
Scaled explained SS	6.781072	Prob. Chi-Square(2)	0.0337

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 05/14/11 Time: 17:37
Sample: 1 40
Included observations: 40

$$\hat{e}_i^2 = \alpha_1 + \alpha_2 x_i + \alpha_3 x_i^2 + v_i$$

$$\chi^2 = N \times R^2 = 40 \times 0.18888 = 7.555$$

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-2908.783	8100.109	-0.359104	0.7216
INCOME	291.7457	915.8462	0.318553	0.7519
INCOME^2	11.16529	25.30953	0.441150	0.6617

R-squared	0.188877	Mean dependent var	7612.629
Adjusted R-squared	0.145032	S.D. dependent var	10873.10
S.E. of regression	10053.75	Akaike info criterion	21.34132

<과제>

8.8

8.16

Eviews output을 출력하고,

출력물의 빈 여백에 간단하게 답을 적으시오.

※ 참고: 필요한 data는 WILEY 교과서 홈페이지에 있음

http://principlesofeconometrics.com/