12주차 강의자료 및 과제 : 이분산(교과서 282-304쪽 참고)

1. 이분산 존재시 최소제곱추정량 b_2를 β_2의 최우수선형불편추정량(BLUE)이라고 할 수 없다.

- 최소제곱추정법(또는 통상적 최소제곱추정법; OLS)에서는 모든 관찰값에 대해 분산이 동일하다는 전제하에 추정량(또는 계수)을 산출하였다. 이때 최소제곱 추정량으로 산출한 계수(b_2)는 확률변수이자, β_2에 대해 선형불편 추정량이다. 뿐만 아니라 최소분산을 갖는다. 이때 우리는 최소제곱추정량(b_2)를 β_2의 최우수선형불편추정량(BLUE)이라고 한다(가우스-마코프 정리 참조).

  (12-1) $SR3(\text{ or } MR3): \ var(e_i) = \sigma^2$

·식(12-1)에서 e_i는 평균이 0이고, 동일한 분산을 가지는 무작위오차항이다.

- 그런데 모든 관찰 값에 대해 분산이 동일하지 않은 경우도 있다. 우리는 이를 이분산이 존재한다고 정의한다. 이 경우 단순회귀모형(또는 다중회귀모형)의 기본가정(SR3 또는 MR3)을 위배하게 된다. 따라서 최소제곱추정량(b_2)는 β_2의 선형불편 추정량일지라도 최소분산을 갖는 추정량은 아니다. 즉, 보다 적은 분산을 갖는 선형불편추정량을 다시 찾아야 한다.

  (12-2) $\cdot SR3(\text{or } MR3): \ var(e_i) = \sigma_i^2 = h(x_i)$

·식(12-2)에서 e_i는 평균이 0이고, x_i가 변동함에 따라 분산이 변동하는 무작위오차항이다..

· 동분산일 경우 b_2에 대한 최소제곱추정량의 분산은 식(12-3)과 같다. 그러나 이분산일 경우 b_2에 대한 최소제곱추정량의 분산은 식(12-4)와 같다.

(12-3) $var(b_2) = \dfrac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ (12-4) $var(b_2) = \dfrac{\sigma_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

·식(12-3)과 식(12-4)의 분산은 전혀 다르다.

- 만일 이분산이 있는 경우에도 최소제곱추정법으로 계수를 산출하고, 분산과 표준오차를 추정해서 신뢰구간을 추정하거나 가설검정을 실시하면 어떻게 될까? 답은 오류를 범하게 된다.
  · 표준오차($se(b_2)$)를 크게 계산함으로써 신뢰구간이 더 넓어지게 된다. 또한 t값이 더 작아진다. 따라서 귀무가설을 수락할 확률이 높아 진다. 즉, 부정확한 신뢰구간 추정이나 부정확한 검정통계량 값을 피해야 한다.

- 따라서 일정방법에 의해 관측치에 대해 가중치를 줌으로써 이분산을 동분산으로 바꾼 후 유효한(efficient) 분산과 표준오차를 산출해서 올바른 신뢰구간을 추정하거나 가설검정을 실시해야 한다. 우리는 이와 같은 방법을 일반최소제곱추정법(generalized least squares method)라한다.

2. 이분산 유형은 비례적 이분산성과 분할적 이분산성으로 구분된다. 따라서 두 유형에 따라 이분산을 동분산으로 바꾸는 절차는 차이가 있다. 하지만 동분산으로 바꾸게 되면 유효한 (efficient) 분산과 표준오차를 산출해서 올바른 신뢰구간을 추정하거나 가설검정을 실시하는 절차는 동일하다.
   두 유형에 따라 이분산을 동분산으로 바꾸는 절차는 다음과 같다.

2.1 비례적 이분산성이 있는 경우 일반화된 최소제곱추정법

- 식 (12-5)의 표본회귀모형에서 x_i가 변동함에 따라 잔차의 분산이 비례적으로 변동한다.

(12-5) $y_i = b_1 + b_2 x_i + \hat{e_i}, \quad var(\hat{e_i}) = \hat{\sigma_i^2} = \hat{\sigma^2} x_i$

- 이 경우 이분산을 동분산으로 변화시키는 절차는 다음과 같다. 식(12-5)의 양변에 $\frac{1}{\sqrt{x_i}}$ 곱하면 식(12-6)과 같다. 다음은 $y_i^* = \frac{y_i}{\sqrt{x_i}}$, $x_{i1}^* = \frac{1}{\sqrt{x_i}}$, $x_{i2}^* = \frac{x_i}{\sqrt{x_i}}$, $e_i^* = \frac{\hat{e_i}}{\sqrt{x_i}}$로 치환하면 식(12-7)이 된다.

(12-6) $\frac{y_i}{\sqrt{x_i}} = \frac{b_1}{\sqrt{x_i}} + b_2 \frac{x_i}{\sqrt{x_i}} + \frac{\hat{e_i}}{\sqrt{x_i}}$

(12-7) $y_i^* = b_1 x_{i1}^* + b_2 x_{i2}^* + e_i^*, \quad E(e_i^*) = 0, \quad var(e_i^*) = var(\frac{e_i}{\sqrt{x_i}}) = \frac{1}{x_i} var(e_i) = \frac{1}{x_i} \hat{\sigma^2} x_i = \hat{\sigma^2}$

- 식(12-7)의 경우 변형된 잔차항의 평균은 0이며, 또한 동분산 조건을 만족한다.

- 동분산 조건을 만족하는 식(12-7)에서 잔차(e_i^*)의 제곱을 합계한 값을 최소로 하는, 즉 최소제곱추정량 산출 방법을 적용하면 식(12-8)과 같이 b_1, b_2가 산출된다. 이때 b_2는 β_2의 선형불편 추정량이며 동시에 최소분산을 갖는 추정량으로 가우스-마코프 정리의 조건을 충족한다. 또한 식(12-9)를 기초로 해서 추정한 분산과 표준오차를 이용해서 신뢰구간을 추정하거나 가설검정을 실시한다.

(12-8) $\hat{y_i^*} = b_1 x_{i1}^* + b_2 x_{i2}^*$

- 한편 식(12-8)의 양변에 $\sqrt{x_i}$ 곱하면 우리에게 익숙한 표본회귀식(12-9)가 유도된다. 우리는 이를 이용하여 탄력성을 산출할 수 있으며, 예측도 할 수 있다.

(12-9) $\hat{y_i} = b_1 + b_2 x_i$

- 다중회귀모형의 경우 비례적 이분산성이 있는 경우 일반화된최소제곱추정법을 정리하면 다음과 같다.

·첫째 단계 : 다중회귀모형에서 어떤 변수가 이분산성에 비례적인가를 결정한다. 그런데 다중회귀모형에서 어떤 변수인지를 사전에 알기가 어렵다. 따라서 우리는 White검정법[1]에 의해 이분산성이 있는가를 검정하자.

ⅰ. 가설설정 : $H_0: var(e_i)=\sigma^2,\ H_1: var(e_i)=\sigma_i^2$

ⅱ. 최소제곱추정법으로 추정한후 잔차를 구한다.

$$y_i = b_1 + b_2 x_{i1} + b_3 x_{i2} + b_4 x_{i3} + ... + b_k x_{ik-1} + \hat{e_i}$$

ⅲ. 최소제곱추정법으로 보조적 회귀식을 추정(auxiliary regression)하고, 결정계수(R^2)를 산출한다.

$$\hat{e_i^2} = \delta_1 + \delta_2 x_{i1} + \delta_3 x_{i2} + \delta_4 x_{i3} + ... + \delta_k x_{ik-1} + \hat{\nu_i}$$

$$H_0 : \delta_2 = \delta_3 = \delta_4 = .. = \delta_k = 0$$

ⅳ. 카이제곱 검정통계량을 구한 후 귀무가설 기각여부를 판별한다. 이때 α는 유의수준, $k-1$은 독립변수 개수이다. 만일 검정통계량이 임계치보다 클 경우 귀무가설을 기각된다. 이 경우 α유의수준하에서 이분산성이 있다고 추론한다,

$$\chi^2 = n \times R^2 \geq \chi^{2_{(\alpha, k-1)}}$$

·둘째 단계 : 이분산성이 있다면 잔차($\hat{e_i}$)의 분산($\hat{\sigma^2}$)을 구한다. 그리고 $\frac{1}{\hat{\sigma_i}}$를 종속변수, 독립변수, 잔차에 가중평균해서 새로운 종속변수와 독립변수로 치환해서 다중회귀모형을 새롭게 재정비한다. 이때 가중평균된 잔차항의 평균은 0이다. 또한 동분산 조건을 만족한다.

$$y_i^* = b_1 x_{i1}^* + b_2 x_{i2}^* + b_3 x_{i3}^* + b_4 x_{i4}^* + e_i^*, \quad E(e_i^*) = 0, \quad var(e_i^*) = var(\frac{e_i}{\hat{\sigma_i}}) = \frac{1}{\hat{\sigma_i^2}} var(e_i) = 1$$

·세째 단계 : 새로운 종속변수와 설명변수들을 가지는 변환된 모형에 대해 최소제곱추정법을 적용한다. 이때 b_i는 β_i의 선형불편 추정량이며 동시에 최소분산을 갖는 추정량으로 가우스-마코프 정리의 조건을 충족한다. 또한 새로운 다중회귀모형 추정결과를 기초로 해서 신뢰구간을 추정하거나 가설검정을 실시한다.

---

1) 화이트 검정법 중 이해를 쉽게 하기 위해서 교차항 없는 경우에 한정한다.

2.2. 분할적 이분산성이 있는 경우 일반화된 최소제곱추정법

- 미국 사회는 개인별로 시간당 임금수준이 매우 다양하다. 이와 같이 개인별 임금 격차를 요인별로 규명하기 위하여 다중회귀모형을 설정하면 식(12-10)과 같다.

(12-10) $WAGE_i = b_1 + b_2 EDUC_i + b_3 \mathrm{EXP}ER_i + b_4 METRO_i + \hat{e_i}$

- 식(12-10)에 따르면 개인별 시간당 임금수준($WAGE_i$)은 개인별 교육수준($EDUC_i$)이 1년 더 많을 경우 b_2달러, 경력기간($\mathrm{EXP}ER_i$)이 1년 더 길수록 b_3달러, 그리고 모의변수인 거주지역($METRO_i$; 대도시지역 =1, 농촌지역=0)이 대도시인 경우 시간당 임금이 b_4달러 더 많은 것을 알 수 있다.

- 그런데 우리는 식(12-10)에서 다음과 같은 의문이 있게 된다. 대도시지역에 거주하는 경우 잔차의 분산이 농촌지역에 거주하는 경우보다 더 크지 않을까? 왜냐하면 대도시지역의 경우 다양한 형태의 직업이 있기 때문에 잔차의 분산이 농촌지역보다 훨씬 더 클 수 있기 때문이다.

- 지금부터 이와 같은 의문에 대해 진위 여부를 검정하자. 또 두 지역간에 잔차의 분산이 다르다면, 이분산 문제를 어떻게 처리해야 할까? 이 문제를 해결하는 방법에 대해 알아보자. 순서는 다음과 같다.

· 첫째, 표본($n = M + R$)을 2개 부표본(대도시지역과 농촌지역)으로 구분하자. 이때 대도시지역 관측치 수는 M개, 농촌지역 괸측치 수는 R개다.

· 둘째, 대도시지역 부표본과 농촌지역 부표본 관측지 자료에 대해 최소제곱추정법으로 각각의 회귀식과 잔차를 추정하자.

M= 대도시지역인 경우 $WAGE_{M,i} = b_{M1} + b_{M2} EDUC_{Mi} + b_{M3} \mathrm{EXP}ER_{Mi} + \widehat{e_{M,i}}$

R= 농촌지역인 경우 $WAGE_{R,i} = b_{R1} + b_{R2} EDUC_{Ri} + b_{R3} \mathrm{EXP}ER_{Ri} + \widehat{e_{R,i}}$

· 세째, 골드펠트-퀀트 검정(Goldfeld-Quandt test)을 실시한다. 대도시지역의 잔차($\widehat{e_{M,i}}$)를 이용하여 분산($\widehat{\sigma_M^2}$)을 구한다. 또 농촌지역의 잔차($\widehat{e_{R,i}}$)를 이용하여 분산($\widehat{\sigma_R^2}$)을 구한다. 이어서 식(12-11)과 같이 F-통계량 값이 임계값보다 클 경우 이분산이 존재한다고 추론할 수 있다.

(12-11) $F = \dfrac{\widehat{\sigma_M^2}}{\widehat{\sigma_R^2}} \geq F_{(\alpha, M-3, R-3)}$

넷째, 이분산이 존재할 경우 동분산으로 전환시키는 절차는 다음과 같다. 대도시지역의 관측치의 경우 식(12-10)의 양변에 $\dfrac{1}{\widehat{\sigma_M}}$를 곱한다. 농촌지역의 경우에는 $\dfrac{1}{\widehat{\sigma_R}}$를 곱한다. 식(12-12)는 새로운 종속변수와 설명변수들을 가지는 변환된 다중회귀모형이다.

(12-12) $$\frac{WAGE_i}{\hat{\sigma}_i} = \frac{b_1}{\hat{\sigma}_i} + b_2\frac{EDUC_i}{\hat{\sigma}_i} + b_3\frac{\mathrm{EXP}ER_i}{\hat{\sigma}_i} + b_4\frac{METRO_i}{\hat{\sigma}_i} + \frac{\hat{e}_i}{\hat{\sigma}_i}$$

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

식(12-12)와 같이 $\frac{1}{\hat{\sigma}_i}$를 종속변수, 독립변수, 잔차에 가중평균해서 새로운 종속변수와 독립변수로 일반화시키면 식(12-13)과 같다. 이때 가중평균된 잔차항의 평균은 0이다. 또한 동분산 조건을 만족한다. 관측치 수는 n개다.

(12-13) $$y_i^* = b_1x_{i1}^* + b_2x_{i2}^* + b_3x_{i3}^* + b_4x_{i4}^* + e_i^*, \quad E(e_i^*) = 0, \quad var(e_i^*) = var(\frac{e_i}{\hat{\sigma}_i}) = \frac{1}{\hat{\sigma}_i^2}var(e_i) = 1$$

·다섯째, 새로운 종속변수와 설명변수들을 가지는 변환된 모형에 대해 최소제곱추정법을 적용한다. 이때 b_i는 β_i의 선형불편 추정량이며 동시에 최소분산을 갖는 추정량으로 가우스-마코프 정리의 조건을 충족한다. 또한 식(12-13)을 기초로 해서 추정한 분산과 표준오차를 이용해서 신뢰구간을 추정하거나 가설검정을 실시한다.

12주차 과제

1. 이분산의 의미를 설명하고, 이분산이 존재할 가능성이 있는 자료의 예를 설명하시오.
2. 브레쉬-페이건 검정에 대해 예를 들어 설명을 하시오.