5강. 보조정보를 이용한 추정

◈ 담당교수 : 이기재 교수

■ 주요용어

용어	해설
주변수(study variable)	표본조사를 통해 각 조사단위로부터 직접적으로 알아내고자 하는 관심변수
보조변수(auxiliary variable)	주변수 이외 조사단위에 관한 정보로 일반적으로 주변수와 연관성이 높은 변수일 경우 유용성이 높음
비추정(ratio estimation)	표본 조사단위로부터 주변수 외에 보조변수의 값도 얻을 수 있을 때 주변수와 보조변수의 비(比)를 이용하는 추정방법
상대효율(relative efficiency : RE)	두 가지 서로 다른 추정량들의 효율을 비교하기 위해 사용되는 측도로 서로 다른 추정량의 분산 비(比)를 뜻함
회귀추정(regression estimation)	두 변수 x와 y사이의 직선 관계식을 가정하여 구하는 추정량으로 회귀분석에서 두 변수 사이의 회귀식을 구하는 것과 유사함

■ 실습하기

- 교재 96쪽 아파트 매매가격 변동비 추정
  * 비추정에 대한 추정값, 표준오차,
    신뢰구간 작성

- 교재 105쪽 심화문제 1-3번
  * 비추정에 의한 모총계 추정,
    표준오차, 신뢰구간 작성

■ 연습문제

1. n개 표본에서의 각 단위들의 주변수 뿐만 아니라 보조변수 값도 함께 구하여 추정에 활용하는 방법으로 비추정법과 회귀추정법이 있다.
주변수와 보조변수가 원점을 지나는 직선관계를 지닐 경우 ( )이 적절한 반면, 원점을 지나지 않는 직선관계일 경우에는 ( )이 적절하다.

정답 및 해설 : 비추정, 회귀추정

2. 두 변수간의 비인 $R=\frac{\tau_y}{\tau_x}$ 에 대한 추정량은 ( )이고, 이 추정량에 대한 분산식은 ( )이다.

정답 및 해설 :

$$r=\sum_{i=1}^{n} y_i / \sum_{i=1}^{n} x_i = \bar{y}/\bar{x}\ ,\quad \hat{V}(r)=\frac{N-n}{N}\frac{1}{n}\frac{1}{\mu_x^2}\frac{\sum_{1}^{n}(y_i-rx_i)^2}{n-1}$$

3. 어느 표본조사에서 보조변수 와 주변수 에 대한 산점도를 그려보니 원점을 지나지 않는 직선관계이고, 두 변수 사이의 상관계수는 -0.85인 것으로 나타났다. 이 경우에 가장 효율적인 추정법으로 생각되는 것은?

① 표본평균을 이용한 추정법
② 비추정법
③ 회귀추정법
④ 총계 추정법

정답 : ③

해설 : 보조변수 x와 주변수 y사이에 선형관계를 가정할 수 있지만 원점을 지나지는 않는 경우에 효율적인 추정법은 회귀추정법이다. 주어진 문제에서 두 변수는 원점을 지나지 않고, 상관계수가 -0.85이므로 직선관계가 뚜렷하다고 할 수 있다. 비추정법은 두 변수 사이의 산점도를 그렸을 때 원점을 통과하는 선형관계를 가정할 수 있을 때 효율적이다.

4. 단순임의표본에 대하여 보조변수 x와 주변수 y를 조사해서, 비추정량을 이용하여 모평균과 모총계를 추정하고자 한다. 여기서, r_x과 μ_x는 각각 보조변수 x에 대한 모집단 총계와 평균으로 사전에 알려져 있다. 다음 중 옳지 <u>않은</u> 것은?

① 모평균에 대한 비추정량은 $\hat{\mu}_y=\mu_x\frac{\bar{y}}{\bar{x}}$ 이다.

② 모총계에 대한 비추정량은 $\hat{\tau}_y = \tau_x \frac{\bar{y}}{\bar{x}}$이다.

③ x와 y의 산점도가 원점을 통과하지 않는 직선 관계일 때는 회귀추정량을 이용한다.

x와 y의 상관계수가 0 근처의 값이면 비추정량이 표본평균에 비해서 효과적이다.

정답 : ④

해설 : 비추정량이 표본평균에 비해서 효과적인 경우는 와 의 상관계수가 1/2보다 큰 경우이다.

5. 단순임의표본을 추출하여 보조변수 x와 주변수 y를 조사하고, 회귀추정량을 이용하여 모평균과 모총계를 추정하고자 한다. 다음의 설명 중 적절하지 않은 것은?

① 모평균에 대한 회귀추정량은 $\hat{\mu}_{yL} = \bar{y} + b(\mu_x - \bar{x})$이다.

② 회귀추정량 $\hat{\mu}_{yL}$의 분산은 $\hat{V}(\hat{\mu}_{yL}) = \frac{N-n}{N}\frac{MSE}{n}$이다. 단, *MSE*: 회귀분석의 평균제곱오차이다.

③ 모총계에 대한 $\hat{\tau}_{yL} = \tau_x \cdot b$회귀추정량은이다.

④ 보조변수 x와 주변수y의 상관계수의 절대값이 1에 가까울 때 표본평균에 비해 효율적이다.

정답 : ③

해설 : 총계에 대한 회귀추정량은 $\hat{\tau}_{yL} = N \cdot \hat{\mu}_{yL} = N \cdot [\bar{y} + b(\mu_x - \bar{x})]$ 이다.

■ 정리하기

1. n개 표본에서의 각 단위들의 주변수 $y_1, y_2, \cdots, y_n$뿐만 아니라 보조변수 $x_1, x_2, \cdots, x_n$값도 함께 구하여 추정

활용하는 방법으로 **비추정법**과 **회귀추정법**이 있다.

2. 주변수 y와 보조변수 x의 비인 $R = \frac{\mu_y}{\mu_x}$의 추정량과 r과 그 분산의 추정량은 다음과 같다. ,

$$r = \frac{\sum_i^n y_i}{\sum_i^n x_i} = \frac{\bar{y}}{\bar{x}}, \quad \hat{V}(r) = \frac{N-n}{N}\frac{1}{n}\frac{1}{\mu^2}\frac{\sum_i^n (y_i - rx_i)^2}{n-1}$$

한편, 분산 추정식을 다르게 표현하면 일반적인 통계 소프트웨어를 이용하여 다음과 같이 간편하게 계산 할 수다.

$$\hat{V}(r) = \frac{N-n}{N}\frac{1}{n}\frac{1}{\mu^2}(s_y^2 + r^2 s_x^2 - 2r\hat{\rho} s_x s_y)$$

여기서, $\hat{\rho}$은 x와 y의 표본상관계수로 $\hat{\rho} = \frac{s_{xy}}{s_x s_y}$이다.

3. **비추정량을 이용하여 모총계를 추정하는 경우** 추정량과 분산의 추정량은 다음 식과 같다.

$$\hat{\tau}_y = r \cdot \tau_x = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}$$

$$\hat{V}(\hat{\tau}_y) = \hat{V}(r \cdot \tau_x) = \tau_x^2 \hat{V}(r) = N^2 \cdot \frac{N-n}{N}\frac{1}{n}\frac{\sum_1^n (y_i - rx_i)^2}{n-1}$$

4. **주변수와 보조변수가 원점을 지나는 직선관계를 지닐 경우 비추정**이 적절한 반면, **원점을 지나지 않는 직선관계일 경우는**

회귀추정이 적절하다. 모평균에 대한 회귀추정량과 분산추정량은 다음과 같다.

$$\widehat{\mu_{yL}} = \bar{y} - b(\mu_x - \bar{x})$$

$$\hat{V}(\widehat{\mu_{yL}}) = \frac{N-n}{N}\frac{1}{n}\frac{1}{n-2}\left[\sum_{i=1}^n (y_i - \bar{y})^2 - b^2 \sum_{i=1}^n (x_i - \bar{x})^2\right] = \frac{N-n}{N}\frac{MSE}{n}$$

여기서 $b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$이다.

5. 추정량 E_1의 E_2에 대한 상대효율은 $RE\left(\frac{E_1}{E_2}\right) = \frac{V(E_2)}{V(E_1)}$로 정의되며, 이 값이 1보다 클수록, E_1이 E_2에 비해

더 효율적인 추정량이라고 한다.

■ 참고문헌

- 이계오, 박진우, 이기재, 표본조사론, 한국방송통대학교출판부, 2013. 제1장
- 통계청 홈페이지 : http://www.nso.go.kr