5강 보조 정보를 이용한 추정

정보통계학과 이기재교수

학/습/목/차

개요

- 표본조사 이론의 두 가지 관심사

1	대표성과 경제성을 고려한 좋은 표본 추출
2	주어진 표본 정보를 잘 활용하는 효율적인 추정법

- 좋은 표본조사 추출
  - 다양한 추출방법 활용
  - 추출 과정에서 보조정보를 활용하는 추출법 : 층화추출법

개요

- 효율적인 추정법

1	단순 추정량 : 관심변수만을 이용하는 추정량 예 $\hat{\mu} = \bar{y} \ , \ \hat{\tau} = N\hat{\mu}$
2	보조변수를 활용한 추정량 : 비추정량, 회귀추정량

- 보조변수 활용의 예

  ▶ 어느 도시의 아파트 가격동향조사

  주변수 : 특정 아파트의 이번 달 가격
  보조변수 : 특정 아파트의 기준월 가격

  ▶ 쌀 생산량 조사

  주변수 : 표본경지의 쌀 생산량, Y
  보조변수 : 표본경지의 벼 재배면적, X

학/습/목/차

1. 개요
2. 비추정
3. 회귀추정
4. 엑셀을 활용한 실습

비(ratio)의 추정

- 모집단
  - 주변수 : $y_1, y_2, \ \dots \ , y_N$
  - 보조변수 : $x_1, x_2, \ \dots \ , x_N$
  - 두 변수간의 비(ratio) : $R = \dfrac{\tau_y}{\tau_x} = \dfrac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i} = \dfrac{\mu_y}{\mu_x}$
- 표본 : N개 중 n개를 단순임의추출
  - 주변수 : $y_1, y_2, \ \dots \ , y_n$
  - 보조변수 : $x_1, x_2, \ \dots \ , x_n$

비(ratio)의 추정

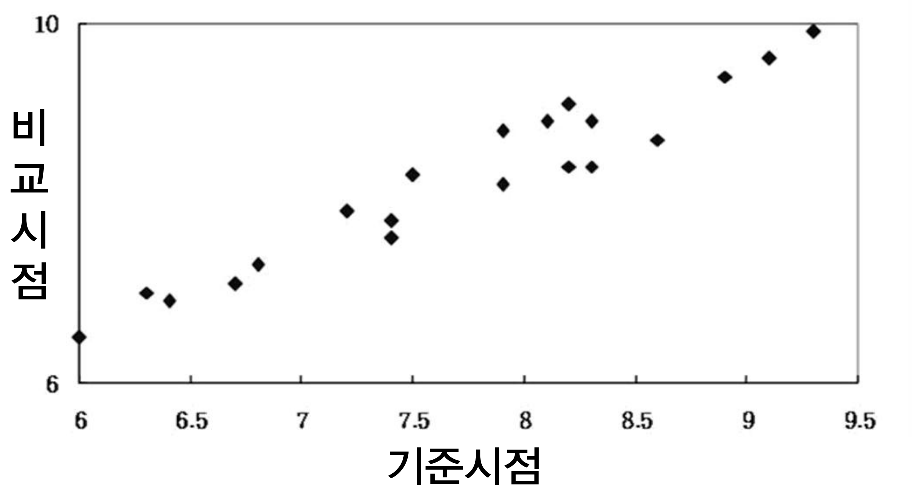
- 비(R)의 추정량
  - $r = \sum_{i=1}^{n} y_i / \sum_{i=1}^{n} x_i = \frac{\overline{y}}{\overline{x}}$
- 비추정량의 분산 추정량
  - $\hat{V}(r) = \frac{N-n}{N} \frac{1}{n} \frac{1}{\mu_x^2} \frac{\sum_{i=1}^{n} (y_i - r x_i)^2}{n-1}$
  - 모평균(μ_x)을 모를 때 추정치 ($\overline{x}$) 을 구하여 대치함
- 모수 R에 대한 $100(1-\alpha)\%$신뢰구간
  - $r \pm z_{\alpha/2} \sqrt{\hat{V}(r)}$

비(ratio)의 추정

예제 3-1

▶ 도시 주택가격의 변화율 추정

- N=1,000 , n=20 호의 단순임의추출
- 주변수(y) : 현재 시점의 조사가격,
  보조변수(x) : 기준 시점의 조사가격



〈그림 3-1〉 표본 데이터의 산점도

비(ratio)의 추정

예제 3-1

▶ 도시 주택가격의 변화율 추정

- 변화율(R)의 추정값 : $r = \dfrac{\text{조사년도 표본주택들의 가격합계}}{\text{기준년도 표본주택들의 가격합계}}$

$$= \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{164.7}{154.5} \fallingdotseq 1.07$$

- 추정분산값 : $\hat{V}(r) = \dfrac{N-n}{N}\dfrac{1}{n}\dfrac{1}{(\bar{x})^2}\dfrac{\sum_{i=1}^{n}(y_i - rx_i)^2}{n-1}$

$$= \frac{1,000-20}{1,000} \cdot \frac{1}{20} \cdot \frac{1}{7.725^2} \cdot \frac{1.2844}{20-1}$$
$$= 0.0000555$$

- R에 대한 95% 신뢰구간 : $r \pm z_{\alpha/2}\sqrt{\hat{V}(r)} \leftrightarrow 1.07 \pm 0.015$

비(ratio)의 추정

예제 3-1

▶ 도시 주택가격의 변화율 추정

〈참고〉 분산추정량의 또 다른 표현법

- 통계소프트웨어 활용한 계산

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

여기서, $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \qquad s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$

$$s_{xy}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \quad \hat{\rho} = \frac{s_{xy}}{s_x s_y}$$

비(ratio)의 추정

예제 3-1

▶ 도시 주택가격의 변화율 추정

〈참고〉 분산추정량의 또 다른 표현법

- 통계소프트웨어의 계산 결과 활용

구분	n	평균	표준편차	상관계수
x	20	7.725	0.947	$\hat{\rho}=0.966$
y	20	8.235	0.957	

$$\hat{V}(r)=\frac{1,000-20}{1,000}\frac{1}{20}\frac{1}{7.725^2}(0.957^2+1.07^2\times 0.947^2-2\times 1.07$$
$$\times 0.966\times 0.947\times 0.957)=0.0000568$$

모총계의 추정

- 모총계와 비(比)의 관계
  - ▶ $R = \dfrac{\tau_y}{\tau_x} \Rightarrow \tau_y = R \bullet \tau_x$
- 비추정량을 이용한 모총계의 추정
  - ▶ $\hat{\tau}_y = r \bullet \tau_x = \dfrac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} \bullet \tau_x$
  - ▶ τ_x 는 일반적으로 조사 전에 미리 알고 있는 경우가 많음
- 분산추정량
  - ▶ $\hat{V}(\hat{\tau}_y) = \tau_x^2 \hat{V}(r) = N^2 \bullet \dfrac{N-n}{N} \dfrac{1}{n} \dfrac{\sum_{i=1}^{n} (y_i - r x_i)^2}{n-1}$

모총계의 추정

예제 3-2

▶ 2010년 서울시의 인구추정

- N=25개구, n=7개 구를 단순임의추출, 보조변수 : 2005년의 인구수
- 비추정값 : $r = \dfrac{\sum_{i=1}^{7} y_i}{\sum_{i=1}^{7} x_i} = \dfrac{2,922,139}{2,886,034} = 1.0125$
- 2005년 인구추정값 : $\hat{\tau}_y = r\tau_x = 1.0125 \times 9,820,171 = 9,943,024$
- 분산추정값 : $\hat{V}(\hat{\tau}_y) = N^2 \cdot \dfrac{N-n}{N}\dfrac{1}{n}\dfrac{\sum_{i=1}^{n}(y_i - rx_i)^2}{n-1}$

$$= 25^2 \cdot \frac{25-7}{25}\frac{1}{7}\frac{4,297,113,521.4}{6}$$

$$= 4.60405 \times 10^{10}$$

- 95% 신뢰구간 계산 : $\hat{\tau}_y \pm z_{\alpha/2}\sqrt{\hat{V}(\hat{\tau}_y)} \leftrightarrow 9,943,024 \pm 429,141$

모평균의 추정

- 비추정량을 이용한 모평균의 추정
  - $\widehat{\mu_y} = r \bullet \mu_x = \frac{\bar{y}}{\bar{x}} \mu_x$
- 분산추정량
  - $\hat{V}(\widehat{\mu_y}) = \mu_x^2 \hat{V}(r)$

$$= \frac{N-n}{N} \frac{1}{n} \frac{\sum_{i=1}^{n} (y_i - r x_i)^2}{n-1}$$

상대효율 (relative efficiency: RE)

서로 다른 추정량들의 효율을 비교하기 위한 척도

- 추정량 E_1 의 E_2에 대한 상대효율(relative efficiency: RE)
  - $$RE\left(\frac{E_1}{E_2}\right) = \frac{V(E_2)}{V(E_1)}$$
  - 상대효율이 1보다 작을수록 추정량 E_1 은 E_2보다 더 효율적임

    반대로 상대효율이 1보다 클수록 추정량 E_1 은 E_2에 비해 효율이 떨어짐

- 상대효율의 추정값
  - $$\widehat{RE}\left(\frac{E_1}{E_2}\right) = \frac{\hat{V}(E_2)}{\hat{V}(E_1)}$$

상대효율 (relative efficiency: RE)

- 비추정량 $\hat{\mu}_y$ 의 단순추정량 $\bar{y}$에 대한 상대효율 추정

▶ $$\widehat{RE}\left(\frac{\hat{\mu}_y}{\bar{y}}\right) = \frac{\hat{V}(\bar{y})}{\hat{V}(\hat{\mu}_y)}$$
$$= \frac{s_y^2}{s_y^2 + r^2 s_x^2 - 2r\hat{\rho} s_x s_y}$$

- 비추정량이 단순추정량보다 효율적이기 위한 조건

▶ $$s_y^2 + r^2 s_x^2 - 2r\hat{\rho} s_x s_y < s_y^2 \Leftrightarrow r s_x^2 < 2\hat{\rho} s_x s_y$$
$$\Leftrightarrow \hat{\rho} > \frac{1}{2}\frac{r s_x}{s_y} = \frac{1}{2}\frac{s_x/\bar{x}}{s_y/\bar{y}}$$
$$\Leftrightarrow \hat{\rho} > \frac{1}{2}$$

➔ 보조변수와 관심변수 간의 상관계수가 1/2 이상이면

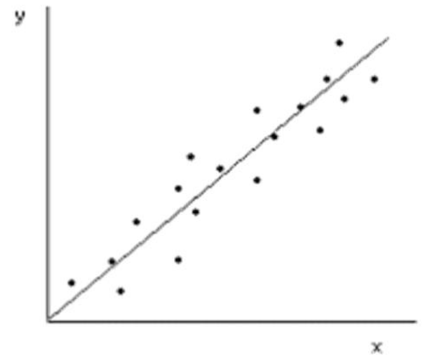
비추정이 더욱 효율적!

학/습/목/차

1. 개요
2. 비추정
3. 회귀추정
4. 엑셀을 활용한 실습

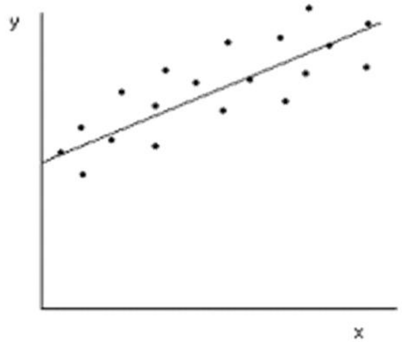
회귀추정과 비추정의 차이

- 비추정
  - ▶ 두 변수 x와 y가 서로 원점을 지나는 직선관계일 때 적용



나무의 나이(x)와 길이(y)

- 회귀추정
  - ▶ 두 변수 x와 y가 원점을 지나지 않는 직선관계일 때 적용



중간고사 성적(x)과 기말고사 성적(y)

모평균 추정

- 회귀 추정량
  - ▶ $\widehat{\mu_{yL}} = \overline{y} + b(\mu_x - \overline{x})$ 여기서, $b = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$

  ➔ b는 $y = \alpha + \beta x + \epsilon$의 회귀모형 하에서 β에 대한 최소제곱추정량임

- 회귀 추정의 추정분산
  - ▶ $$\hat{V}(\widehat{\mu_{yL}}) = \frac{N-n}{N}\frac{1}{n}\frac{1}{n-2}\left\{\sum_{i=1}^{n}(y_i - \overline{y})^2 - b^2\sum_{i=1}^{n}(x_i - \overline{x})^2\right\}$$
    $$= \frac{N-n}{N}\frac{MSE}{n}$$

    (MSE: 회귀분석의 평균제곱오차)

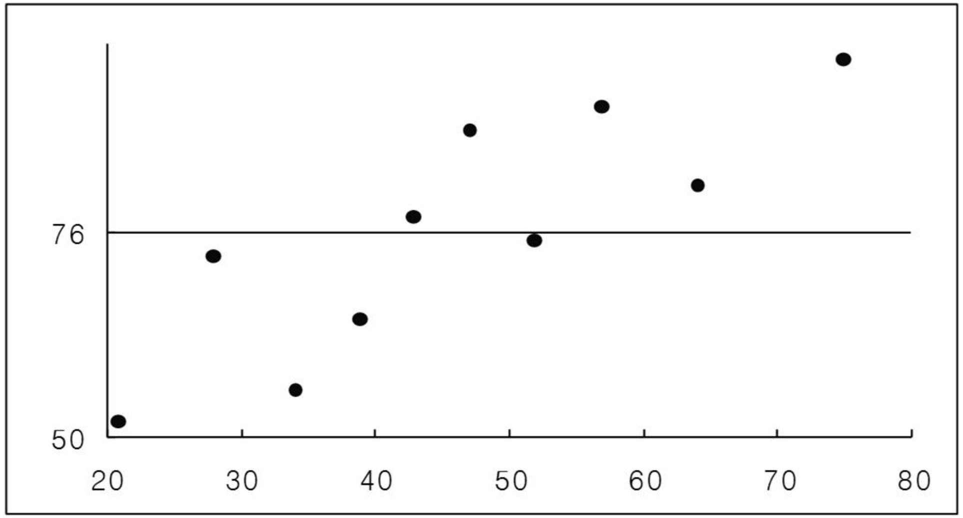
- 모평균에 대한 $100(1-\alpha)\%$ 신뢰구간
  - ▶ $\left(\widehat{\mu_{yL}} - z_{\alpha/2}\sqrt{\hat{V}(\widehat{\mu_{yL}})}\ ,\ \widehat{\mu_{yL}} + z_{\alpha/2}\sqrt{\hat{V}(\widehat{\mu_{yL}})}\right)$

모평균 추정

예제 3-4

▶ 회귀추정 사례

- $N = 486$, $n = 10$, $\mu_x = 52$



- $\bar{y} = 76$, $\bar{x} = 46$, $b = 0.7656$

모평균 추정

예제 3-4

▶ 회귀추정 사례

- 모평균 추정값 : $\widehat{\mu_{yL}} = \bar{y} + b(\mu_x - \bar{x})$
$$= 76 + 0.7656 \times (52 - 46) = 80.59$$

- 추정분산값 : $\hat{V}(\widehat{\mu_{yL}}) = \frac{N-n}{N}\frac{MSE}{n}$
$$= \frac{486-10}{486}\frac{75.75}{10} = 7.42$$

- 모평균에 대한 95% 신뢰구간 :

$$\widehat{\mu_{yL}} \pm z_{\alpha/2}\sqrt{\hat{V}(\widehat{\mu_{yL}})} \leftrightarrow 80.59 \pm 5.45$$

상대효율 비교

- 회귀추정량의 분산식

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

$$\Leftrightarrow \hat{V}(\widehat{\mu_{yL}}) = \frac{N-n}{N}\frac{1}{n}\left\{s_y^2 - b^2 s_x^2\right\} \qquad \left(\leftarrow \frac{1}{n-2} \approx \frac{1}{n-1}\right)$$

$$\Leftrightarrow \hat{V}(\widehat{\mu_{yL}}) = \frac{N-n}{N}\frac{1}{n}s_y^2(1-\hat{\rho}^2) \qquad \left(\leftarrow b = \hat{\rho}s_y/s_x\right)$$

- 회귀추정량의 단순추정량에 대한 상대효율

► $$\widehat{RE}\left(\frac{\widehat{\mu_{yL}}}{\overline{y}}\right) = \frac{s_y^2}{s_y^2(1-\hat{\rho}^2)}$$

$$= \frac{1}{1-\hat{\rho}^2}$$

➔ 주변수와 보조변수 사이의 상관계수(ρ)가 클수록 회귀추정량은 더 효율적!

상대효율 비교

예제 3-5

▶ 상대효율의 계산

- 단순추정 : $\bar{y} = 76$, $s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = 228.4$

$$\hat{V}(\bar{y}) = \frac{N-n}{N}\frac{s^2}{n} = \frac{486-10}{486}\frac{228.4}{10} = 22.37$$

- 회귀추정 : $\widehat{\mu_{yL}} = 80.6$, $\hat{V}(\widehat{\mu_{yL}}) = 7.42$, $\hat{\rho} = 0.84$

- 상대효율 : $\widehat{RE}\left(\dfrac{\widehat{\mu_{yL}}}{\bar{y}}\right) = \dfrac{\hat{V}(\bar{y})}{\hat{V}(\widehat{\mu_{yL}})} = \dfrac{22.37}{7.42} = 3.015$

➡ 즉, 이 예제에서 회귀추정량은 단순추정량보다 3배 가량 높은 요율을 나타냄

학/습/목/차

〈실습하기〉에서 자세히 다룸

이 강의는

강의용 휴대폰(U-KNOU 서비스 휴대폰)으로도

다시 볼 수 있습니다.