유한모집단에서 모형-기반 합성추정치의 예측

신민웅[1)] 김익찬[2)]

요 약

소지역에서 유한모집단의 총계등을 추정하는데 있어서 모형-기반 합성치를 예측한다. 즉, 예측(prediction) 문제로 추정치를 다룬다. 초모집단(super-population) 확률 모형을 세우고 최적의 예측치를 유도한다.

주요 용어 : 유한모집단, 합성추정치, 초모집단 모형

1. 서론

표본조사는 지리적으로 전체지역(whole area)에 대한 추정을 목적으로 하고 있다. 그러나 지방지역(local area)에 대한 통계치를 생산하는 문제도 지방 자치제등으로 중요성이 높아가고 있다. 소지역(small area)은 지리적 지역과 마찬가지로 모집단의 연령-성별-교육 수준에 의한 사회-경제적 분류일 수도 있다. 소지역은 층화의 방법에 의하여 정의되는 모집단의 임의의 부분이 될 수도 있다. 제조업의 예로서, 소지역은 생산품의 한 묶음(batch)일 수도 있다.

명백히, 소지역에 대한 표본크기는 작을 것이고, 어떤 지역은 크기가 영(zero)일수도 있다. 그렇게 작은 표본 크기는 조사 추정치들의 신뢰구간을 매우 크게 만들어 받아들일 수 없게 만든다. 따라서 소지역에 대한 신뢰할만한 추정치들을 얻기 위해서는 특별한 방법들이 필요하다.

직접 추정량은 보통 해당 소지역에서 조사된 자료만을 이용하며 추정된다. 그리고 센서스나 행정자료로부터 획득된 보조 정보를 조사 자료에 추가하여 추정되기도 한다.

합성추정법은 소지역 추정시 소지역을 포함하는 대영역의 정보를 함께 이용하는 방법으로써 소지역과 대영역의 특성 구조가 유사하다는 가정 아래서 이용된다. 합성추정량의 분산은 직접 추정량의 분산에 비해 작으나 앞에서 가정이 성립하지 않을 경우에는 심각한 바이어스(bias)가 발생할 수 있다.

Gonzalez(1973)은 소지역(small area)들이 큰 지역들과 같은 특성을 갖는다는 가정 아래서 합성추정량을 제안하였다. 이것은 큰 지역의 추정량이 소지역의 추정치들을 유도하는데 사용되는 것이다. 합성 추정 방법은 주(state) 수준에서 신체장애 추정을 위하여 US -NCHS(1968)에서 처음으로 사용되었다. 건강 면접조사에서 나이, 성별, 가구의 크기등에 의하여 정의되는 78개의 사후-층에 대한 신체장애의 국가적 비율을 구하였다. 각 주에 대한 신체장애의 합성 추정치를 구하기 위하여 각 주의 알려진 가중 값들을 결합하였다. Laake(1978)은 노르웨이 노동력 조사에서 노르웨이 인구 및 주

1)(449-791) 경기도 용인시 모현면. 한국외국어대학교 정보통계학과, 교수
mwshin@stat.hufs.ac.kr
2)(690-756) 제주도 제주시 아라1동 1 제주대학교 전산통계학과 교수
ickim@cheju.cheju.ac.kr

택조사(1970)를 사용하여 합성추정량의 평균제곱오차를 유도하였다. 합성추정량은 연령과 성별에 의한 사후-층화에 의하여 구하였다. 대규모 가구 조사에서 합성추정치의 효율성은 Levy (1977)의 건강조사, Gonzalez(1973)의 인구 조사, Purcell(1976)의 호주 노동력 조사에서 평가되었다. 그러한 추정량들은 설계-기반(design-based)과 모형-기반(model- based) 의 2가지가 있다.

2. 직접추정량

유한 모집단 P가 N개의 단위로 이루어 졌다고 하자. 즉, $P = \{1,\cdots,N\}$ 으로 N은 알려진 수이고, 단위들은 $1,\cdots,N$ 으로 나타내어진다. P는 크기가 $N_a (a = 1,\cdots,A)$ 인 소지역 P_a로 나누어진다.

즉, $\bigcup_a P_a = P$, $\sum_a N_a = N$ 이다. s_{ag}는 칸 (a, g)에 속하는 표본 s의 부분으로서 크기는 n_{ag}이다. $s_a = \bigcup_g s_{ag}$ 는 소지역 a에 속하는 크기 n_a인 표본의 부 분이다. n_{ag}는 확률변수이다. $s_{0g} = \bigcup_a s_{ag}$이다. 이 논문에서 사용되는 기호는 다음과 같다.

$$\sum_g N_{ag} = N_{a0} = N_a$$
$$\sum_g N_{ag} = N_{0g}$$
$$\sum_a \sum_g N_{ag} = N$$

크기 n의 표본 s가 모집단 P로부터 추출된다. s_{ag}는 칸 (a, g)로부터의 표본이다.

$$n_a = n_{a0} = \sum_g n_{ag}$$
$$n_{0g} = \sum_a n_{ag}$$
$$\sum_a \sum_g n_{ag} = n$$

모평균과 표본평균은 다음과 같다.

$$\bar{Y}_a = \sum_{P_a} y_k / N_a$$

$$\bar{y}_{s_a} = \bar{y}_a = \frac{1}{n_a} \sum_{s_a} y_k \quad \text{(소지역 a에 대한 표본 평균)}$$

$$\bar{y}_{s_{0g}} = \bar{y}_{0g} = \sum y_k / n_{0g} \quad \text{(영역 g에 대한 표본 평균)}$$

$$\bar{y}_{s_{ag}} = \bar{y}_{ag} = \sum y_k / n_{ag} \quad \text{(칸 (a, g)에 대한 표본 평균)}$$

그리고

$$\sum_{A} z_k = \sum_{k \in A} z_k$$

$$\bar{s} = P - s$$

$$\bar{s}_a = P_a - s_a$$

Y_a의 추정량으로 가장 흔히 쓰이는 직접 추정량은 Horvitz-Thompson 추정량으로

$$Y_{a;HT} = \sum_{s_a} \frac{y_k}{\pi_k} \tag{2.1}$$

이다. 이 불편 추정량은 비효율적이지만 다른 추정량들과 비교하는 기준(bench-mark)추정량이다. 여기서, π_k는 k번째 단위가 표본에 포함되는 포함확률(inclusion-probability)이고, y_k는 k번째 단위에 대한 측정치이다. 이 불편 추정량은 비효율적이지만 다른 추정량들과 비교하는 기준(bench-mark)추정량이다.

G 영역에 기반을 둔 사후층화 HT-형 추정량은

$$\begin{aligned} Y^{(G)}_{a;HT/c} &= \sum_g N_{ag} \left(\sum_{s_{ag}} y_k/\pi_k\right)\left(\sum_{s_{ag}} 1/\pi_k\right)^{-1} \\ &= \sum_g (N_{ag}/\, N_{ag;HT})\left(\sum_{s_{ag}} y_k/\pi_k\right) \\ &= \sum_g N_{ag}(\, Y_{ag;HT}/\, N_{ag;HT}) \\ &= \sum_g Y_{ag;Ha/c} \end{aligned} \tag{2.2}$$

단, $Y_{ag;Ha/c}$는 $Y_{ag}(=\sum_{p_{ag}} y_k)$의 Ha'jek 추정량이다. 그리고 $N_{ag;HT} = \sum_{s_{ag}} 1/\pi_k$ 인데, s_{ag}에 속하는 k단위에 관련된 가중값 $1/\pi_k$ 의 합이다.

보조변수 x에 대하여 사후층화 HT-형 비추정량은

$$Y_{a;HT/r}{}^{(G)} = \sum_g X_{ag}(\, Y_{ag;HT}/\, X_{ag;HT}) \tag{2.3}$$

단, $X_{ag} = \sum_{p_{ag}} x_k$.

만일 모든 표본들 S_{ag}가 공집합이 아니라면, 사후층화 추정량 (3.2.2), (3.2.4)는 근사적으로 불편추정량이다. 어떤 칸(cell)이 빈 칸(zero counts)이라면 칸들을 합쳐서 빈 칸을 없애야 한다.

3 . 모형-기반 합성 추정량

모형-기반 아래서 확률변수 y_{agk}가 초모집단 모형 ξ를 따를 때에 우리는 소지역 총계 $\sum_{p_a} y_k = Y_a$ 를 예측하는데 관심이 있다. 초모집단(super-poplation) ξ의 모수는 때때로

보조변수 $x=(x_1, \cdots, x_p)'$ 에 의존한다. 예측(prediction) 이론에서는 표본자료 $(y_{agk}, k\in s)$와 모형 ξ에 포함된 정보를 사용하여 Y_a의 최적 예측량을 구한다
모형

$$y_{agk}=\beta_g+\varepsilon_{agk}$$
$$\varepsilon(\varepsilon_{agk})=0, \quad \varepsilon(\varepsilon^2_{agk})=\sigma^2 \tag{3.1}$$

을 생각하자.
이것은 분산 분석 모형으로 (3.3.25)에서, β_g의 최선의 선형 불편추정량(BLUE)은

$$\mathcal{B}^*_g=\bar{y}_{0g} \tag{3.2}$$

이다.
포함확률 π_k를 고려한 β_g의 가중 추정량은

$$\mathcal{B}_{g\Pi}=(\sum_{s_{og}} y_k/\pi_k)(\sum_{s_{og}} 1/\pi_k)^{-1} \tag{3.3.}$$

이다.
임의의 가중행렬 $Q=Diag(g_k, k=1,\cdots,N)$ 에 대하여, β_g의 Q-가중 추정량은

$$\mathcal{B}_{gQ}=(\sum_{s_{0g}} y_k q_k)(\sum_{s_{0g}} q_k)^{-1} \tag{3.4}$$

이다.
Y_a의 Royall BLU-가중 합성 추정량은 (β_g의 추정치가 $\bar{y}_{og}$이므로)

$$\begin{aligned} \mathcal{T}_a^{*(G)} &= \sum_g\Big[\sum_{sag} y_k+\sum_g\sum_{sag}\mathcal{B}^*_g\Big] \\ &= \sum_g\Big[\sum_{sog} y_k+(N_{ag}-n_{ag})\bar{y}_{0g}\Big] \\ &= \sum_g\Big[n_{ag}(\bar{y}_{cg}-\bar{y}_{0g})+N_{ag}\bar{y}_{0g}\Big] \end{aligned} \tag{3.5}$$

이다. 이식은 Holt, Smith, Tomberlin (1979)의 (3.3)식에서 유도되었다.

G = 1에 대하여

$$\mathcal{T}^*_a=\sum_{sa} y_k+(N_a-n_a)\bar{y}_s \tag{3.6}$$

이다

(3.3.25)에서 $T_a^{*(G)}$의 예측 평균 제곱오차는

$$\begin{aligned}\varepsilon\,(\,T_a^{*(G)}-Y_a)^2 &=\sigma^2\sum_g(N_{ag}-n_{ag})\,(N_{ag}-n_{ag}+n_{0g})/\,n_{0g}\\ &\simeq\sigma^2\sum_g N^2{}_{ag}/\,n_{0g}\ \text{ for } N_{ag}\gg n_{ag}\end{aligned}\tag{3.7}$$

이다.

σ^2의 분산의 추정은

$$\hat{\sigma}^2=\sum_a\sum_g\sum_{k\in s_{ag}}(\,y_k-\bar{y}_{0g})^2/(n-G)\tag{3.8}$$

이다.

Y_a의 일반화 회귀(greg-)합성 예측량은

$$T^{*(G)}{}_{aGR}=\sum_g[\sum_{sog}y_k/\pi_k+(N_{ag}-\sum_{sag}1/\pi_k)\,\bar{y}_{0g}]\tag{3.9}$$

이다.

Samdual(1984)는 다음과 같은 Y_a의 추정량을 제안하였다.

$$\begin{aligned}T_{a,GR(\Pi V)}(X,v) &=\sum_{p_a}\hat{y}_k\div\sum_{s_a}e_k/\pi_k\\ &=T_{a,greg}\end{aligned}\tag{3.10}$$

(3.10)의 첫 번째 항은 Y_a의 사영-합성 추정량으로

$$T_{a,\Pi V}(X,v)=\sum_{p_a}\hat{y}_k\tag{3.11}$$

이다. 두 번째 항은 (3.10)의 설계-바이어스에 대한 수정항(correction term)이다.

4. 일반화된 모형-기반 합성추정량

모집단을 A개의 소지역으로 나누고, 각 소지역을 다시 G개의 영역(층)으로 나누는 경우를 생각하자.

영역 g에서 $y_{g1}, y_{g2}, \cdots y_{gNg}$은 독립이고

$E(y_{gk})=\beta_g x_{gk}$, $V(y_{gk})=\sigma_g^2 v_{gk}$, $cov(y_{gk}, y_{gk'})=0\ \ (k\neq k)$

인 모형을 생각하자. 여기서 , $k=1,2,\cdots,N_{0g}$, $g=1,2,\cdots,G$ 이다. 즉, A개의 소지역과 G개의 영역이 있다고 가정하자, Y_a의 Royall의 BLU-합성 추정량은

$$\hat{T}^{*}_{a}=\sum_{g}\sum_{s_a} y_{gk} +\sum_{g}\hat{\beta}_{g}^{*}\sum_{\bar{s}_a} x_{gk}\ , \qquad \bar{s}_a = P_a - s_a \tag{4.1}$$

Mukhopadhyay(1998)의 예제 3.2에 의하면

$$\hat{\beta}_{g}^{*}=\left\{\sum_{s_{0g}} x_{gk}y_{gk}/v_{gk}\right\}\left\{\sum_{s_{0g}} x^{2}_{gk}/v_{gk}\right\}^{-1} \qquad k=1,2\cdots N_{0g},\quad g=1,2,\cdots,G \tag{4.2}$$

Royall(1970)의 결과를 확장하면

$$E_{\xi}(\hat{T}^{*}_{a}- Y_a)^{2}=\sum_{g}(\sum_{\bar{s}_{ag}} x_{gk})^{2}\sigma^{2}_{g}/(\sum_{s_{0g}} x_{gk}/v_{gk})+\sum_{g}\sigma_{g}^{2}\sum_{\bar{s}_{ag}} v_{gk} \tag{4.3}$$

이다.

참고문헌

1. 캐나다 노동력 조사 방법론(2001) 통계기획국,조사관리과
2. 표본설계(2001) 신민웅, 이상은. 교우사.
3. Small area estimation in survey sampling(1998) Parimal Mukhopadhyay
4. Introduction to small area estimation(2001) JON N.K.Rao. ISI(2001,Korea)
5. Laake, Petter(1979). A predictive approach to subdomain estimation in finite population. Journal of the American Statistical Association, 74. 355-358.