소지역에서 Pseudo-EBLUP 추정

신민웅[1] 백정용[2] 김익찬[3]

요 약

소지역 모형들은 고정된(fixed)효과와 랜덤 효과를 포함하는 일반적 선형 혼합 모형의 특별한 경우로 간주될 수 있다. 소지역 평균이나 총계는 고정된 효과와 랜덤 효과의 일치 결합으로 표현될 수 있다.

블록 대각 공분산 구조를 갖는 선형 혼합모형(mixed model) 아래서 EBLUP은 실제 문제에 있어서 많이 소지역 모형에 응용된다. 설계 가중값(design weight) 들에 의존하고 설계-일치(design consistency) 성질을 만족하는 Pseudo-EBLUP 추정량들은 소지역추정에서 합해지면 (aggregated) 사후-수정(post-adjustment)없이 벤치마킹 성질을 만족한다.

주요용어 : 혼합모형, EBLUP, Pseudo-EBLUP

1.서론

표본조사는 전체 모집단에 대한 추정만 아니라 부모집단에 대한 정보도 얻으려 한다. 표본조사는 지리적으로 전체지역(whole area)에 대한 추정을 목적으로 하고 있으나 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.

소지역 추정에서 작은 표본은 간접 추정량의 필요성을 일으킨다. 간접 추정량으로 효과적으로 표본크기를 증가시키고, 표준오차를 감소시킨다. 간접 추정량으로 합성추정량과 복합추정량이 있다.

합성추정법은 소지역 추정시 소지역을 포함하는 대영역의 정보를 함께 이용하는 방법으로써 소지

1) (449-791) 경기도 용인시 모현면 한국외국어대학교 정보통계학과 교수
mwshin@stat.hufs.ac.kr

2) (449-791) 경기도 용인시 모현면 한국외국어대학교 정보통계학과 박사과정

3) (690-756) 제주도 제주시 아라1동 제주대학교 전산통계학과 교수
ickim@cheju.cheju.ac.kr

역과 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 주택조사(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.

블록 대각 공분산 구조를 갖는 선형 혼합모형(mixed model) 아래서 EBLUP은 실제 문제에 있어서 많이 소지역 모형에 응용된다. 설계 가중값(design weight) 들에 의존하고 설계-일치(design consistency) 성질을 만족하는 pseudo-EBLUP 추정량들은 소지역 추정에서 합쳐지면(aggregated) 사후-수정(post-adjustment)없이 벤치마킹 성질을 만족한다.

소지역 모형들은 고정된(fixed)효과와 랜덤 효과를 포함하는 일반적 선형 혼합 모형의 특별한 경우로 간주될 수 있다. 소지역 평균이나 총계는 고정된 효과와 랜덤 효과의 일치 결합으로 표현될 수 있다.

단순히 행정적의 편리성으로 층화하였을 때, 각 층을 소지역으로 간주하여 선형 혼합모형에서 Pseudo-EBLUP 추정량을 구한다.

2. EBLUP 추정량

최선의 선형 불편 예측량(BLUP)은 모형를 최소화한다. BLUP(best linear unbiased prediction)은 모형에 있는 랜덤 효과들의 분산들에 의존한다. EBLUP 추정량은 분산 모수의 추정량을 대치하므로서 BLUP으로 부터 얻을 수 있다.

설계 가중값들 $w_j(s)$는 Y의 설계-기반 추정량들을 구하는 데 중한 역할을 한다. 이러한 기본 가중값들은 표본 s와 원소 j (j∈s)에 의존한다.

단위 수준 모형 (unit level model)에서 추정량은 설계 가중값들 w_{ij} 이용하지 않는다. w_{ij}는 표본 원소 (i, j)에 대응하는 가중값으로 $j=1,\cdots,n_i$; $i=1,\cdots,m$이다. 결과적으로, 표본설계가 자체 가중이 아닌 한, 설계-일치 추정량이 아니다.

즉, 모든 j 에 대하여 $w_{ij}=w_i$이면 설계-일치 추정량이다.

지역 수준모형(area level model)에서 EBLUP 추정량은 설계 일치가 된다. Prasad와 Rao(1999) 는 조사를 시행하는 사람은 설계 일치, 모형-기반 추정량을 사용하기를 권장하였다. 본 절에서는 설계 가중값들에 의존하고 설계-일치 성질을 만족하는 pseudo-EBLUP 추정량들을 다룬다. 특히 모든 (i, j)에 대하여 $k_{ij}=1$인 등오차 분산들이 성립하는 경우($\sigma_{ij}=\sigma_e^2$)를 생각한다.

행정적인 편리성으로 층화한 후에 층화 일단 집락추출에서 i층(i 지역)에서 j번째 집락(또는 j번째 segment)을 pps 로 추출한다고 하자.

즉,

$$\pi_{ij} = n_i \frac{M_{ij}}{M_{i0}}$$

$$w_{ij} = 1/\pi_{ij} = \frac{M_{i0}}{n_i M_{ij}}$$

모형

$$y_{ij} = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + v_i + e_{ij} \qquad (2.1)$$
$$j = 1, \cdots, n_{ij}, \; i = 1, \cdots, m$$

을 생각하자.

$\overline{w_{ij}} = w_{ij} / \sum_k w_{ik}$ 라 하면

$$\begin{aligned} \bar{y}_{iw} &= \sum_j \overline{w}_{ij} y_{ij} \\ &= \sum_j \overline{w}_{ij} (\beta_0 + \beta_1 \bar{x}_{ij1} + \beta_2 \bar{x}_{iw2} + v_i + \bar{e}_{ij}) \\ &= \beta_0 + \beta_1 \bar{x}_{i1w} + \beta_2 \bar{x}_{iw2} + v_i + \bar{e}_{iw} \end{aligned} \qquad (2.2)$$

여기서,

$$\bar{e}_{iw} = \sum_j \overline{w}_{ij} e_{ij}$$
$$E(\bar{e}_{iw}) = 0$$
$$V(\bar{e}_{iw}) = \sigma_e^2 \sum_j \overline{w}_{ij}^2 = \sigma_e^2 \delta_{iw}$$
$$\bar{x}_{iw} = \sum_j \overline{w}_{ij} x_{ij}$$

3. Pseudo-EBLUP 추정량

먼저, 모수를 β, σ_e^2, 그리고 σ_v^2이 기지라고 가정한다. 그러면,

$$\mu_i = \beta_0 + \overline{X}_{i1}\beta_1 + \overline{X}_{i2}\beta_2 + v_i \qquad (3.1)$$

에서, μ_i의 BLUP추정량은

$$\bar{\mu}_{iw}^H = \beta_0 + \overline{X}_{i1}\beta_1 + \overline{X}_{i2}\beta_2 + r_{iw}(\bar{y}_{iw} - \beta_0 - \bar{x}_{i1w}\beta_1 - \bar{x}_{i2w}\beta_2) \qquad (3.2)$$

이다. 여기서, $r_{iw} = \sigma_v^2/(\sigma_v^2 + \sigma_e^2 \delta_{iw})$이다. 분산 성분들 σ_e^2과 σ_v^2을 REML 방법을 써서 추정한다.

회귀 모수 β를 추정하기 위하여, $(\beta, \sigma_e^2, \sigma_v^2)$이 주어졌을 때, v_i의 BLUP 추정량을 구하면,

$$\bar{v}_{iw}(\beta, \sigma_e^2, \sigma_v^2) = r_{iw}(\bar{y}_{iw} - \beta_0 - \bar{x}_{i1w}\beta_1 - \bar{x}_{i2w}\beta_2) \tag{3.3}$$

이다.

그러면, β에 대한 다음의 설계-가중 추정방정식을 푼다.

$$\sum_i^m \sum_j^{n_i} w_{ij} x_{ij}[y_{ij} - \beta_0 - x_{ij1}\beta_1 - x_{ij2}\beta_2 - \bar{v}_{iw}] = 0 \tag{3.4}$$

여기서, $x_{ij} = (1, x_{ij1}, x_{ij2})^T$이다.
식 (3.4)에서

$$\bar{\beta}_w = [\sum_i \sum_j w_{ij} X_{ij}(X_{ij} - r_{iw}\bar{X}_{iw})^T]^{-1} \times [\sum_i \sum_j w_{ij}(X_{ij} - r_{iw}\bar{X}_{iw}) y_{ij}] \tag{3.5}$$

이다.

σ_e^2과 σ_v^2이 주어졌을 때에, 추정량 $\bar{\beta}_w$은 β에 대한 모형-불편이다. σ_e^2과 σ_v^2을 추정량 $\hat{\sigma}_e^2$과 $\hat{\sigma}_v^2$으로 대치하여, β의 설계-가중 추정량 $\hat{\beta}_w = \hat{\beta}_w(\hat{\sigma}_e^2, \hat{\sigma}_v^2)$를 구할 수 있다. μ_i의 pseudo-EBLUP 추정량은 $(\beta, \sigma_e^2, \sigma_v^2)$을 $(\hat{\beta}_w, \hat{\sigma}_e^2, \hat{\sigma}_v^2)$으로 대치하여 구할 수 있다.
즉,

$$\bar{\mu}_{iw}^H = \hat{\beta}_{0w} + \bar{X}_{i1}\hat{\beta}_{1w} + \bar{X}_{i2}\hat{\beta}_{2w} + r_{iw}(\bar{y}_{iw} - \hat{\beta}_{0w} - \bar{x}_{i1w}\hat{\beta}_{1w} - \bar{x}_{i2w}\hat{\beta}_{2w}) \tag{3.6}$$

이다.

여기서, $\hat{r}_{iw} = \hat{\sigma}_v^2/(\hat{\sigma}_v^2 + \delta_i \hat{\sigma}_e^2)$
설계 가중값(design weight) 들에 의존하고 설계-일치(design consistency) 성질을 만족하는 Pseudo-EBLUP 추정량들은 소지역추정에서 합해지면 (aggregated)되면 사후-수정(post-adjustment)없이 벤치마킹 성질을 만족한다.

즉, $\sum_i N_i \hat{\mu}_{iw}$ 은 직접 조사 회귀추정량 $\hat{Y}_W + (X - \hat{X}_W)^T \hat{\beta}_w$ 와

같다.(N_i는 i번째 지역의 모집단 크기이다.) 여기서,

$\hat{Y}_w = \sum_i w_i \bar{y}_{iw} = \sum_i \sum_j w_{ij} y_{ij}$ 와

$\hat{X}_w = \sum_i w_i \bar{x}_{ij} = \sum_i \sum_j w_{ij} x_{ij}$ 는 각각 Y와 X의 직접 추정량이다. 즉,

$\sum_i N_i \hat{\mu}_{iw}^H = \hat{Y}_w + (X - \hat{X}_w)^T \hat{\beta}_w$

이다.

이와 같이 EBLUP 추정량 $\hat{\mu}_i^H$와 다르게 pseudo-EBLUP 추정량 $\hat{\mu}_{iw}^H$ 는 아무런 수정없이 벤치마킹 성질을 만족한다.

참고문헌

1. Small area estimation (2003). Rao, J.N.K. A John Wiley & Sons,Inc, Publication.
2. You, Y., and Rao, J.N.K.(2002a). A Pseudo-Empirical Best Linear Unbiased prediction Approach to small area estimation using survey weights. Canadian Journal of Statistics, 30,431-439.
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. 캐나다 노동력 조사 방법론(2001) 통계기획국,조사관리과