설계가중치를 이용한 유사 최량선형 비편향 예측

신동윤[1] 신민웅[2]

요 약 : You 와 Rao (2002)는 소지역 추정시 유사 최량선형 비편향 예측에서 설계 가중값을 사용하는 방법을 발전시켰다. 특히 소지역 평균들을 추정하기 위하여 유사-최량선형 비편향 예측 추정량을 제안하였다. 우리는 소지역 추정에서 실용적으로 이용되는 몇 가지 추가적인 성질을 연구하였다.

주요용어 : 혼합모형, EBLUP, Pseudo-EBLUP

1.서론

소지역 추정에서 작은 표본은 간접 추정량의 필요성을 일으킨다. 간접(indirect) 추정량으로 효과적으로 표본크기를 증가시키고, 표준오차를 감소시킨다.간접추정량은 추정하려는 소지역과 관련있는 지역으로 부터의 정보를 이용한다. 간접 추정량으로 합성추정량과 복합추정량이 있다.

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.

2.내포(nested) 오차 선형회귀 모형

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

단위 수준 모형 (unit level model)에서 추정량은 설계 가중값들 w_{ij} 를 이용하지 않는다. 여기서, w_{ij} 는 표본원소인 i지역에 j번째 원소에 대응하는 가중값으로 $j=1,\cdots,n_i$; $i=1,\cdots,m$ 이다. 그러면, 가중값을 이용하지 않으므로 표본설계가 자체 가중이 아닌 한,

1) 경기도 용인시 모현면 한국외국어대학교 정보통계학과 박사과정

2) (449-791) 경기도 용인시 모현면 한국외국어대학교 정보통계학과 교수
mwshin@stat.hufs.ac.kr

설계-일치 추정량이 아니다. 만약에 모든 j 에 대하여 $w_{ij}=w_i$ 이면 설계-일치 추정량이다.

모집단 모형

$$y_{ij}=\beta_0+\beta_1 x_{ij1}+\beta_2 x_{ij2}+v_i+e_{ij}$$
$$j=1,\cdots,N_i\ ,i=1,\cdots,m$$

을 생각한다. 여기서, y_{ij}는 i번째 소지역의 j번째 단위(unit)변수, x_{ij1}과 x_{ij2}는 보조변수이고, N_i는 i번째 소지역의 모집단 단위의 개수이다. 랜덤 효과 v_i는 $iid\ N(0,\sigma_v^2)$, 단위오차 e_{ij}는 $iid\ N(0,\sigma_e^2)$ 이라고 가정한다. i번째 소지역의 평균 $\overline{Y_i}$의 조사값은

$$\theta_i=\overline{X_i}\beta+v_i$$

이다.

i번째 ED(Enumeration District)에서 j번째 가구(또는 j번째 segment)을 srs 로 추출한다고 하자. 즉, ED는 소지역으로 가구는 모집단 단위(unit)로 간주한다. srs 로 표본을 추출하므로

$$\pi_{ij}=\frac{n_i}{N_i}$$

$$\tilde{w}_{ij}=1/\pi_{ij}=\frac{N_i}{n_i}$$

그러면,

$$w_{ij}=\frac{\tilde{w}_{ij}}{\sum_{j=1}^{n_i}\tilde{w}_{ij}}=\frac{N_i/n_i}{\sum N_i/n_i}=\frac{1}{n_i}$$

표본 모형

$$y_{ij}=\beta_0+\beta_1 x_{ij1}+\beta_2 x_{ij2}+v_i+e_{ij} \tag{2.1}$$

$$j=1,\cdots,n_i,\quad i=1,\cdots,m$$

을 생각하자. n_i는 i번째 소지역의 표본크기이다.

$$\begin{aligned}\bar{y}_{iw}=\sum_{j=1}^{n_i}w_{ij}y_{ij}&=\sum_j\frac{1}{n_i}y_{ij}\\&=\sum_j\frac{1}{n_i}(\beta_0+\beta_1\bar{x}_{ij1}+\beta_2\bar{x}_{iw2}+v_i+\bar{e}_{ij})\\&=\beta_0+\beta_1\bar{x}_{i1w}+\beta_2\bar{x}_{iw2}+v_i+\bar{e}_{iw}\end{aligned} \tag{2.2}$$

이고, $\overline{X_i} = (x_{i1} + \cdots + x_{iN_i})/N_i$ 이다.

여기서,

$$\bar{e}_{iw} = \sum w_{ij} e_{ij} = \sum_j \frac{1}{n_i} e_{ij}$$

$$E(\bar{e}_{iw}) = 0$$

$$V(\bar{e}_{iw}) = \sigma^2_e \sum_j (\frac{1}{n_i})^2 = \sigma^2_e / n_i$$

$$\bar{x}_{iw} = \sum w_{ij} x_{ij} = \sum_j \frac{1}{n_i} x_{ij}$$

3. Pseudo-EBLUP 추정량

먼저, 모수를 β, σ^2_e, 그리고 σ^2_v이 기지라고 가정한다. 그러면,

$$\Theta_i = \beta_0 + \overline{X}_{i1}\beta_1 + \overline{X}_{i2}\beta_2 + v_i \tag{3.1}$$

에서, Θ_i 의 BLUP추정량은

$$\hat{\Theta}_{iw} = \beta_0 + \overline{X}_{i1}\beta_1 + \overline{X}_{i2}\beta_2 + r_{iw}(\bar{y}_{iw} - \beta_0 - \bar{x}_{i1w}\beta_1 - \bar{x}_{i2w}\beta_2) \tag{3.2}$$

이다. 여기서, $r_{iw} = \sigma^2_v/(\sigma^2_v + \sigma^2_e \delta_{iw})$이다. 분산 성분들 σ^2_e과 σ^2_v을 REMIL 방법을 써서 추정한다.

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

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

이다.

그러면, β에 대한 다음의 설계-가중 추정방정식을 푼다.

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

식 (3.4)에서

$$\bar{\beta}_w = [\sum_i \sum_j \frac{N_i}{n_i} x_{ij}(x_{ij} - r_{iw}\bar{x}_{iw})^T]^{-1}[\sum_i \sum_j \frac{N_i}{n_i}(x_{ij} - r_{iw}\bar{x}_{iw}) y_{ij}] \tag{3.5}$$

이다.

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

비례확률 추출을 한다면,

$$\hat{\beta}_w=[\sum_i\sum_j x_{ij}(x_{ij}-r_{iw}\bar{x}_{iw})^T]^{-1}[\sum_i\sum_j(x_{ij}-r_{iw}\bar{x}_{iw})y_{ij}] \quad (3.6)$$

이다.

Neyman 할당을 하면, 식 (3.5)의 n_i는

$n_i=n\frac{N_iS_i}{\Sigma N_iS_i}$ 이다.

여기서, S_i는 i번째 소지역의 모 표준편차이다.

σ_e^2과 σ_v^2이 주어졌을 때에, 추정량 $\hat{\beta}_w$은 β에 대한 모형-불편추정량이다. 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. θ_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,

$$\hat{\theta}_{iw}=\hat{\beta}_{0w}+\bar{X}_{i1}\hat{\beta}_{1w}+\bar{X}_{i2}\hat{\beta}_{2w}+r_{iw}(\bar{y}_{iw}-\hat{\beta}_{0w}-\bar{x}_{i1w}\hat{\beta}_{1w}-\bar{x}_{i2w}\hat{\beta}_{2w}) \quad (3.7)$$

이다.

여기서, $\hat{r}_{iw}=\hat{\sigma}_v^2/(\hat{\sigma}_v^2+\delta_i\hat{\sigma}_e^2)$이다.

참고문헌

1. George E. Battese, Rachel M. Harter and Wayne A. Fuller.(1988). "An error components. model for predition of county crop areas using survey and satellite data." Journal of the American Statistical Association. March 1988, Vol 83. pp.28-36.

2. Introduction to small area estimation(2001) JON N.K.Rao. ISI(2001,Korea).

3. Small area estimation in survey sampling(1998) Parimal Mukhopadhyay.

4. Small area estimation (2003). Rao, J.N.K. A John Wiley & Sons,Inc, Publication.

5. 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.