J. of Basic Sciences, Cheju Nat'l Univ.
20(1), 83~88, 2007

기초과학연구 제주대학교
20(1), 83~88, 2007

Pseudo-EBLUP 방법에 의한 소지역 추정

김 익 찬
제주대학교 전산통계학과

요 약

소지역 모형들은 고정된(fixed)효과와 랜덤 효과를 포함하는 일반적 선형 혼합 모형의 특수한 경우로 간주될 수 있고 이 경우 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 소지역 모형에 많이 응용된다. 설계 가중값(design weight)들에 의존하고 설계-일치(design consistency)성질을 만족하는 Pseudo-EBLUP 추정량들은 소지역내의 표본크기 n_i를 적당히 증가시켜 동일한 오차분산이 되는 특수한 경우 사후-보정(post-adjustment) 없이 벤치마킹 성질을 만족한다. 본 연구에서는 복합추정량을 이용하여 서귀포시 밀감생산량을 모의 추정하였다.

1. 서 론

표본조사의 목적에 따라 다소 차이가 있지만, 표본설계 시 가장 중요한 변수의 정도, 혹은 모집단 추정치의 정도에 맞춰 표본의 크기를 결정하므로 소지역에 대한 표본크기는 작을 수밖에 없으며, 어떤 도메인은 할당된 표본이 전혀 없을 수도 있다. 따라서 이러한 소지역에 대한 일반적인 추정량들(직접 추정량)은 변동(variation)이 커서 분산의 추정치를 과대 왜곡시키고 결과적으로 신뢰구간을 매우 크게 만들어 추정치의 의미를 상실시킨다. 그럼에도 불구하고 소지역 추정에 대한 다양한 욕구를 만족시키기 위해 많은 연구가 있어 왔다.

직접 추정량은 일반적으로 해당 소지역에서 조사된 자료만을 이용하여 추정된다. 그러나 가능한 경우 센서스나 행정 자료로부터 보조 정보를 얻어 이를 조사 자료에 추가함으로서 추정치의 정도를 높이는 간접추정방법이 활용되면서 1970년대 중반부터 소지역 추정법에 대한 연구가 활발히 진행되었다.

간접 추정량은 크게 합성추정량과 복합추정량으로 구분되는데, 합성추정량은 소지역 추정 시 소지역을 포함하는 대영역의 정보를 함께 이용하는 방법으로서 소지역과 대영역의 특성구조가 유사하다는 가정 아래서 이용된다. 합성추정량의 분산은 직접 추정량의 분산에 비해 작으나 소지역의 특성 구조가 대지역의 특성구조와 유사하지 않을

경우에는 심각한 바이어스(bias)가 발생할 수 있다.

블록 대각 공분산 구조를 갖는 선형 혼합 모형(mixed model) 아래서 경험적 최량선형비편향추정량(Empirical Best Linear Unbiased Prediction; 이하 EBLUP)은 실제 문제에 있어서 소지역 모형에서 많이 응용된다. 설계 가중값(design weight)들에 의존하고 설계-일치(design consistency)성질을 만족하는 Pseudo-EBLUP 추정량들은 소지역 추정에서 합쳐지면(aggregated) 사후-보정(post-adjustment)없이 벤치마킹 성질을 만족한다.

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

단순히 행정적인 편리성으로 층화하였을 때, 각 층을 소지역으로 간주하여 선형 혼합 모형에서 Pseudo-EBLUP 추정량을 구한다.

2. EBLUP 추정량

소지역 모형이란 추정하려는 변수와 상관관계가 높은 보조변수들을 활용하여 도메인 간의 변의에 영향을 미치는 임의의 도메인 특성을 빌려오는 모형을 의미한다. 따라서 소지역 모형은 고정된 효과와 랜덤인 효과를 포함하는 일반적인 혼합모형의 특수한 경우로 간주될 수 있다.

다음의 일반적 선형혼합모형을 생각하자.

$$\mathbf{y}=\mathbf{X}\beta+\mathbf{Z}\mathbf{v}+\mathbf{e} \tag{2.1}$$

$\mathbf{y}=(\mathrm{n}\times 1)\,vector$, $\mathbf{X}=(\mathrm{n}\times \mathrm{p})$,

$\mathbf{Z}=(\mathrm{n}\times \mathrm{h})$ 랜덤행렬

$\mathbf{v}\sim(\mathbf{0},\mathbf{G})$, $\mathrm{e}\sim(\mathbf{0},\mathbf{R})$

단 $\mathbf{G},\mathbf{R}$은 분산모수 $\delta=(\delta_1,\ \delta_2,\ \cdots,\ \delta_g)^T$에 의존하는 공분산 행렬이다.

이제 1차 결합 $\mu=\mathbf{l}^{\mathbf{T}}\beta+\mathbf{m}^{\mathbf{T}}\mathbf{v}$ ($\mathbf{l}$과 $\mathbf{m}$은 특수한 *vector*)를 추정하려고 한다. δ가 알려졌을때 μ의 BLUP 추정량

$$\tilde{\mu}=t(\delta,\mathbf{y})=\mathbf{l}^{\mathbf{T}}\tilde{\beta}+\mathbf{m}^{\mathbf{T}}\tilde{\mathbf{v}} \tag{2.2}$$

단

$$\tilde{\beta}=\tilde{\beta}(\delta)=(\mathbf{X}^{\mathbf{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{V}^{-1}\mathbf{y}$$

$$\tilde{\mathbf{v}}=\tilde{\mathbf{v}}(\delta)=\mathbf{G}\mathbf{Z}^{\mathbf{T}}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\tilde{\beta})$$

여기서 모수 δ를 추정량 $\hat{\delta}=\hat{\delta}(y)$로 둘 때 $t(\hat{\delta},\mathbf{y})$를 EBLUP 추정량이라고 한다. 즉 BLUP은 모형에 있는 랜덤 효과들의 분산에 의존하며 EBLUP 추정량은 분산 모수의 추정량으로 대치함으로서 BLUP로부터 얻을 수 있다.

한편 설계 가중값들 $w_i(s)$는 설계-기반 추정량들을 구하는데 중요한 역할을 한다. 이 기본 가중값 들은 표본 s와 원소 $j(j\in s)$에 의존한다.

단위보조 데이터 $x_{ij}=(x_{ij1},\cdots,x_{ijp})^T$를 활용할 수 있는 단위수준모형(unit level model)을 생각하자.

$$y_{ij}=x_{ij}^{T}\beta+v_i+e_{ij}:$$

$$j=1,\cdots,N_i,\ \ i=1,\cdots,m \tag{2.3}$$

이 모형에서 EBLUP 추정량 $\hat{\mu}_i^H$는 표본 원소 (i,j) ; $j=1,\cdots,n_i,\ i=1,\cdots,m$에 대응하는 설계 가중값 w_{ij}를 활용할 수 없다. 즉 표본설계가 지역 내에서 자체 가중이 아니면 설계-일치 추정량이 아니다. 이때 모든 j에 대해서 $w_{ij}=w_i$로 두면 설계-일치 추정량이 된다.

다음절에서 설계 가중값에 의지하고 설계-일치 성질을 만족하는 Pseudo-EBLUP 추정량을 다룬다. 특히 모든 $(i,\ j)$에 대하여 $k_{ij}=1$인 등 오차 분산이 성립하는 경우 $(\sigma_{ij}=\sigma_e^2)$를 생각한다.

3. Pseudo-EBLUP 추정량

모수 β, σ_e^2, 그리고 σ_v^2이 기지라고 가정하자. 그러면,

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

에서, μ_i의 BLUP추정량은

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

이다. 여기서, $r_{iw}=\sigma_v^2/(\sigma_v^2+\sigma_e^2\delta_{iw})$이다. 분산 성분들 σ_e^2과 σ_v^2을 제한적 최대우도추정법(restricted maximum likelihood estimation ; REML)을 써서 추정한다.

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

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

이다.

β에 대한 다음의 설계-가중 추정방정식을 푼다.

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

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

식 (3.4)에서

$$\tilde{\beta}_w(\sigma_e^2,\sigma_v^2)= \left[\sum_i\sum_j w_{ij}X_{ij}(X_{ij}-r_{iw}\overline{X}_{iw})^T\right]^{-1}\times \left[\sum_i\sum_j w_{ij}(X_{ij}-r_{iw}\overline{X}_{iw})y_{ij}\right] \qquad (3.5)$$

σ_e^2과 σ_v^2이 주어졌을 때에, 추정량 $\tilde{\beta}_w$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 $\hat{\beta}_w=\tilde{\beta}_w(\hat{\sigma}_e^{\,2},\hat{\sigma}_v^{\,2})$를 구한다. 그러면 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 대치함으로써 구해진다.

즉,

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

이다.

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

추정량 $\hat{\mu}_{iw}^{H}$는 그 가중치가 i 상에서 합해져서 즉, $\sum_j w_{ij} = w_{i.} = N_i$가 알려지고 $x_{ij1} = 1$ 인 단위수준모형의 중간 항을 포함하면 자동적으로 벤치마킹 성질을 만족한다.

전체 모집단 크기 Y와 X의 직접 추정량은 $\hat{Y}_w = \sum_i w_{i.}\, \bar{y}_{iw} = \sum_i \sum_j w_{ij} y_{ij}$,

$\hat{X}_w = \sum_i w_i \cdot \bar{\mathbf{x}}_{iw} = \sum_i \sum_j w_{ij} x_{ij}$이고,

$$\sum_i N_i \hat{\mu}_{iw}^{H} = \hat{Y}_w + (X - \hat{X}_w)^T \cdot \hat{\beta}_w \qquad (3.7)$$

이 성립한다.(N_i는 i번째 지역의 모집단 크기)

왜냐하면 β와 $\tilde{v}_{iw}(\beta, \sigma_v^2, \sigma_e^2)$대신에 $\hat{\beta}_w$와 $\hat{v}_{iw} = \hat{v}_{iw}(\hat{\beta}_w, \hat{\sigma}_v^{2.} \hat{\sigma}_e^{2})$으로 대치하고 $x_{ij1} = 1$로 두면 (3.4)에서

$$\sum_i \sum_j w_{ij}(y_{ij} - \mathbf{x}_{ij}^{T} \hat{\beta}_w - \hat{v}_{iw}) = 0 \quad \text{또는}$$

$$\sum_i N_i \hat{v}_{iw} = \hat{Y}_w - \hat{X}_w^{T} \hat{\beta}_w \qquad (3.8)$$

이 성립한다. 그러면

$\hat{\mu}_{iw}^{H} = \bar{X}_i^{T} \hat{\beta}_w + \hat{v}_{iw}$이고, 식 (3.6)을 이용하면,

$$\sum_i N_i \hat{\mu}_{iw}^{H} = X^T \hat{\beta}_w + \sum_i N_i \hat{v}_{iw}$$
$$= \hat{Y}_w + (X - \hat{X}_w)^T \hat{\beta}_w \qquad (3.9)$$

즉, Pseudo-EBLUP 추정량 $\hat{\mu}_{iw}^{H}$는 EBLUP 추정량 $\hat{\mu}_i^{H}$와는 달리 아무런 수정없이 벤치마킹 성질을 만족한다.

4. 응용과 결론

동별	농가수	재배면적 (ha)	표본수	면적 (a)	수확량 (kg)
송산, 보목	479	167	2	33 67	13125 15750
천지, 남성	80	31	1	50	15000
정방, 중앙	0	0			
효돈	1054	364	4	33 50 100 133	15000 15750 31875 45000
영천	1224	689	4	27 30 67 133	11250 13125 16875 41250
동홍	482	224	2	23 50	9375 15750
서홍	615	289	3	33 67 167	13125 26250 56250
대륜	1150	532	4	40 43 100 30	15000 15750 31875 11250
대천	1564	635	5	233 20 50 60 33	75000 7500 18750 24375 12000
중문	1265	817	5	40 33 67 27 30	15000 12000 22500 13125 11250
예래	1237	681	4	43 37 100 23	15750 12000 24375 7500
계	9450	4429	34		

위의 표는 (3.9) 식의 적용에 의한 실례를 확인하기 이전에 모의 실험을 통해서 2005년도 서귀포시 조생밀감 예상량을 구하기 위한 추정자료이다. 실 농가주 와의 면접을 통해 재배하는 조생면적과 예상 생산량을 물었다. 서귀포시 12개동 중 정방동과 중앙동은 실제 농가가 없어 표본크기가 영이 되는 경우가 발생하였는데 대여력(borrowing strength)에 의해 추정 가능한 부분이므로 2개동을 제외한 10개 동에 속한 총 밀감 생산 농가수 N_i는 80에서 1565농가에서 총 모의표본 n=34 농가를 추출하였고 각 동에서 추출한 농가는 1에서 5가구이다. 종래에 추정해오던 엽과비 방식보다는 생산농가가 직접적으로 추정하는 면적 대 예상생산량이 직접적인 상관관계가 높음에 착안하여 이를 보조변수로 하고 각 지역 i내의 모든 표본단위에 대하여 기본적인 가중치 $w_{ij}=w_{i.}=\frac{N_i}{n_i}$를 적용하였다.

GLS(Generalized Least Squares) 방식을 이용하여 β를 추정한 회귀모형은

$$\widehat{y_{ij}}=-2132.9+369x_{ij}$$

이고 분산의 추정치 $\widehat{\sigma^2}=18536620$ 이었다. 2005년도 서귀포시의 조생밀감 총 예상 생산량은

$$\sum_{i=1}^{m} N_i\ \hat{\mu}_{iw}{}^{H}=\ \hat{Y}_w+(X-\ \hat{X}_w)^T\hat{\beta}_w$$

$=190,053$ ton 이다. 본 연구는 직접조사에 활용할 수 있는 Pseudo-EBLUP 추정량의 방법 중 밀감생산자의 직접적인 추정을 보조자료로 삼는 것이 생산량 추정에 타당한지를 확인하고자 하였으며 분산성분의 확인을 위한 추가적인 연구와 실제 자료의 확보가 요청된다.

<표> 2005년도 제주도 서귀포시 조생밀감 생산량추정

동별	농가수 (N_i)	n_i	Pseudo-EBLUP	s.e	$\hat{Y}$
송산, 보목	479	2	16.3	1.37	7807.7
천지, 남성	80	1	16.3	1.15	1304.0
효돈	1054	4	27.0	0.34	28458.0
영천	1224	4	21.6	0.98	26438.4
동홍	482	2	11.3	1.15	5446.6
서홍	615	3	30.7	1.08	18880.5
대륜	1150	4	17.6	0.93	42384.4
대천	1564	5	27.1	0.66	19393.6
중문	1565	5	12.4	1.54	19406.0
예래	1237	4	16.6	1.31	20534.2
계	9450	34			190053.4

참고문헌

[1] 서귀포시청(2005) 감귤재배실태 일제조사, 서귀포시

[2] 최기헌. 최지영.(2004) 회귀모형에 의한 소지역 추정, Journa of the Korean Data Analysis Society Vol 6. No 6. December, 1715 - 1723

[3] Battese, G.E. Harter, R.M and Fuller, W.A.(1988) An error components model for prediction of county crop area using survey and satellite data, Journal of American Statistical Association, Vol.83, NO.401, 28-36

[4] Gonzalez, M.E.(1973). Use and evaluation of synthetic estimators, Proceedings of the Social Science Section, American Statistical Association, 33-36

[5] Parimal Mukhopadhyay(1998). Small area estimation in Survey Sampling, Narosa Publishing House, London

[6] Rao, J.N.K. (2003). Small area estimation, A John Wiley & Sons, Inc, Publication, New York

[7] You, Y. and Rao, J.N.K.(2002). A Pseudo-Empirical Best Linear Unbiased Prediction approach to small area estimation using survey weights, Canadian Journal of Statistics, Vol. 30, 431-439.