碩士學位論文

소지역 추정량의 활용방안에 관한 연구

濟州大學校 大學院

電 算 統 計 學 科

魯 永 玉

2009 年 8 月

소지역 추정량의 활용방안에 관한 연구

指導敎授 金 益 贊

魯 永 玉

이 論文을 理學 碩士學位 論文으로 提出함

2009 年 8 月

魯永玉의 理學 碩士學位 論文을 認准함

審査委員長 김 철 수 ㊞

委 員 김 익 찬 ㊞

委 員 이 윤 정 ㊞

濟州大學校 大學院

2009 年 8 月

A Study of Application for Estimator of Small Area

Young-Ock Noh
(Supervised by Professor Ik-Chan Kim)

A thesis submitted in partial fulfillment of the requirement for the degree of Master of Science

2009 . 8

This thesis has been examined and approved.

Chul-Soo Kim

Ik-Chan Kim

Yoon-Jung Rhee

August 2009

Department of Computer Science and Statistics
GRADUATE SCHOOL
JEJU NATIONAL UNIVERSITY

목 차

표 목 차

요 약

소지역추정(small area estimation)이란 배정된 표본크기가 작은 소지역이나 성별, 연령, 교육수준, 소득수준 등과 같은 변수의 특성으로 분류된 소영역(small domain)에 대한 통계를 생산하는데 이용되는 추정방법이다. 이러한 소지역은 통계를 추정하기에는 표본이 극히 적어지게 되고, 이에 따라 소지역에 대한 조사추정치의 신뢰구간을 매우 크게 만듦으로서, 신뢰성 있는 소지역 통계를 얻기 어렵게 된다. 따라서, 이를 극복하고 소지역에 대한 신뢰할만한 추정치를 얻기 위해서는 특별한 방법이 필요하고, 이러한 방법을 연구하는 것이 소지역추정이다.

본 논문에서는 소지역 추정방법으로 직접 추정량, 합성추정량, 복합추정량에 대해 소개하고, 논문에서 주로 다루고자 하는 Pseudo-EBLUP 추정량을 알아보기 위해 그 기반이 되는 내포오차선형회귀모형과 EBLUP 추정량을 설명하고 최종적으로 Pseudo-EBLUP 추정량을 알아보았다. 본 논문에서는 소지역 추정량 활용 사례로 제주지역 서귀포시 감귤 생산량을 추정, 비교 분석하고 제안한 추정량의 효율적 활용을 위하여 표본조사시에 필요한 추가자료 확보를 제안함으로서 그 활용의 뜻을 높이고자 하였다.

Ⅰ 서론

소지역추정(small area estimation)이란 배정된 표본크기가 작은 소지역이나 성별, 연령, 교육수준, 소득수준 등과 같은 변수의 특성으로 분류된 소영역(small domain)에 대한 통계를 생산하는데 이용되는 추정방법이다.

표본조사의 목적에 따라 다소 차이가 있지만, 표본설계시 가장 중요한 변수의 정도, 혹은 모집단 추정치의 정도에 맞춰 크기를 결정하므로 소지역에 대한 표본크기는 작을 수밖에 없으며, 어떤 소지역은 할당된 표본이 전혀 없을 수도 있다.

예를 들어, 소지역을 경제활동인구조사를 비롯한 대부분의 정부통계에서의 시군구 등으로 볼 수도 있다. 또한, 소지역을 제조업의 예로서 부품소재산업의 생산품의 한 묶음(batch)으로 볼 수도 있다. 소지역을 층화의 방법에 의해 정의되는 모집단의 임의의 부분으로 볼 수도 있다. 위의 각 경우 소지역에 대한 통계 추정의 경우 대부분의 통계를 생산하기 위한 표본설계는 소지역을 포함하는 대영역의 통계를 생산할 목적으로 설계되기 때문에, 소지역에 대한 통계를 추정하기에는 표본이 극히 적어지게 되고, 현실적인 비용을 고려할 때 소지역에 대한 신뢰성 있는 통계를 얻기 위해 새로운 표본조사를 실시하는 것이 불가능한 경우가 발생할 수 있게 된다. 그렇게 되면, 소지역에 대한 조사추정치의 신뢰구간을 매우 크게 만듦으로서, 신뢰성 있는 소지역 통계를 얻기 어렵게 된다. 따라서, 이를 극복하고 소지역에 대한 신뢰할만한 추정치를 얻기 위해서 특별한 방법이 필요하고, 이러한 방법을 연구하는 것이 소지역추정이다.

본 논문에서는 소지역 추정 방법으로 직접추정량을 언급하고, 합성추정량, 복합추정량에 대하여 정리하였다. 내포오차선형회귀 추정량과 EBLUP 추정량, Pseudo-EBLUP 추정량에 대해서 정리하였다.

직접 추정량은 일반적으로 해당 소지역에서 조사된 자료만을 이용하여 추정된다. 직접추정량은 비효율적이지만 다른 추정량들과 비교되는 기준이 된다.

가능한 경우 센서스나 행정 자료로부터 보조 정보를 얻어 이를 조사 자료에 추가함으로서 추정치의 정도를 높이는 간접추정방법이 활용되면서 1970년대 중반

부터 소지역추정법에 대한 연구가 활발히 진행되었다.

간접 추정량은 크게 합성추정량과 복합추정량으로 구분되는데, 합성추정량은 소지역추정시 소지역을 포함하는 대지역의 정보를 함께 이용하는 방법으로 소지역과 대지역의 특성구조가 유사하다는 가정 아래서 이용된다. Gonzalez(1973)는 하나의 마트로시카(러시아 인형)안에 같은 모양과 형태를 지닌 인형들이 그 크기만 달리하여 들어있는 것처럼 소지역들이 대지역들과 같은 특성을 갖는다는 가정 아래서 합성추정량을 제안하였다. 이 방법은 소지역 포함하는 대지역의 추정량으로 소지역의 추정치들을 유도하는 방법이다. 복합추정량은 합성추정량의 한계를 보완하기 위해서 사용되어지는 추정량이다. 합성추정량의 경우 소지역이 대지역과 같은 특성을 갖는다는 가정이 있어야 하는데 이러한 가정에서 벗어날 경우 설계 바이어스가 매우 크게 된다는 문제가 발생하게 되는데 이를 보완하기 위해, 직접추정량과 합성 추정량에 적절한 가중치를 사용하여 결합한 형태의 추정량이 된다.

내포오차선형회귀 추정량은 관심변수에 대한 추정을 위해 보조변수를 사용하고 그 형태를 모형으로 구성하여 추정하는 방법을 의미한다. 블록 대각 공분산 구조를 갖는 선형 혼합모형(mixed model)아래서 경험적 최량선형비편향추정량(empirical best linear unbiased prediction : 이하 EBLUP)은 실제 문제에 있어서 소지역 모형에서 많이 응용된다. 그러나 EBLUP가 설계일치 추정량이 되지 못하므로 하여 응용의 한계가 있음에 반하여 설제 가중값(design weight)들에 의존하고 설계-일치(design consistency)성질을 만족하는 Pseudo-EBLUP 추정량은 소지역 추정에서 이 두 성질이 합쳐지면(aggregated) 사후-보정(post-adjustment) 없이 벤치마킹 성질을 만족한다.

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

이를 위한 활용으로 서귀포지역 조생밀감 생산량 추정에 사용하였다. 조사자료는 2008년도 서귀포 농업기술원에서 수집한 자료를 본 연구에 활용하였다. 서귀포지역의 각 동을 소지역으로 설정하고 조사된 자료에서 관심변수로 1주당 생산량을 정하고, 1주당 열매수와, 10a당 주수를 보조변수로 하여 모집단 자료에서 각 동의 조생밀감 재배면적을 활용하여 서귀포 지역내의 총 생산량을 추정하였다.

Ⅱ 소지역 추정량

1. 직접 추정량

직접 추정량은 오직 소지역에서 얻은 조사 자료를 이용하여 추정하는 방법이다

1) 총계 추정을 위한 추정량

총계(total) 추정을 위한 가장 단순한 추정량으로 다음 식을 사용한다.

$$\hat{Y}_a = \sum_{i=s_a} w_i y_i \qquad 1\text{-}(1)$$

여기서, n_a : 소지역인 a지역에서 추출된 표본

w_i : i번째 관측치의 가중치(일반적으로 추출률의 역수)

$\hat{Y}_a$는 불편 추정량이지만 표본크기에 따라 변동량이 크게 나타날 수 있다는 단점이 있다.

2) 사후 층화 추정량

사후 층화 추정량(post-stratified estimator)은 각 부분집단의 크기는 알고 있으나 부분집단별 목록이 없는 경우에 사용되며, 우선 임의의 표본을 뽑은 후 추정단계에서 층별 자료를 이용하여 추정량의 정도를 향상시키는 방법이다.

따라서 소지역인 a의 크기 (N_a)가 알려진 경우, 사후 층화 추정량을 이용하며, 이 추정량의 수식은 다음과 같다.

$$\hat{Y}_{pos.a} = N_a \frac{\sum_{i=s_a} w_i y_i}{\sum_{i=s_a} w_i} = N_a \frac{\hat{Y}_a}{\hat{N}_a} = N_a \overline{\hat{y}_a} \qquad 1\text{-}(2)$$

이 방법은 식 1-(1) 보다 안정된 추정량을 제공하지만 조사설계가 복잡한 경우 비추정편의(ratio estimation bias)가 발생할 수 있다는 단점을 가지고 있고 분산은 다음 식처럼 구해진다.

3) Horvitz-Thompson 추정량

Horvitz-Thompson 추정량은 총계(total)의 추정량으로 가장 많이 쓰이는 직접 추정량으로 수식은 다음과 같다.

$$\hat{Y}_{HT.a} = \sum_{i=1}^{N} \frac{y_i}{\pi_i} \qquad 1\text{-}(3)$$

$U = \{1, 2, \cdots, N\}$, $Y = \sum_{i}^{N} y_i$: Y의 모집단 총합

$A \subset U$, $\pi_i = \Pr[i \in A] : U_i$ 단위(원소 i)가 표본에 포함되는 확률

이 불편추정량은 비효율적이지만 다른 추정량들과 비교되는 기준(bench-marking)추정량이 된다.

2. 간접추정량

1) 합성 추정량(Synthetic Estimator)

간접추정의 한 방법으로서 합성추정을 다음과 같이 정의할 수 있다. 소규모의 지역은 대규모의 지역과 같은 특성을 가진다는 가정하에서 대규모의 지역으로부터 얻어진 불편추정치를 소규모 지역에 대한 추정치로 보며, 이 추정치를 합성추정이라고 한다.

합성추정량은 소지역 표본 추정치 중 전통적이고, 가장 흔히 쓰이는 방법이다. 이 방법은 추정 기법이 간단하고 원표본설계의 응용이 가능하며 추정치를 구하려는 소지역과 대지역의 정보를 빌려서 추정치의 정도를 높이는 방법이다.

합성추정량은 해당 소지역과 유사한 특성을 지니고 있는 대지역들의 정보를 이용하여 추정되기 때문에 직접추정량에 비하여 그 추정오차가 현저하게 줄어들 수 있지만 해당 소지역이 속해 있는 하위그룹내의 소지역들에 대한 정보가 해당 소지역의 것과 동질적이라는 가정이 성립하지 않게 되면 편향(bias)이 커지게 될 가능성이 있다.

대지역이 i개의 소지역을 가지고 있다고 하고, 대지역의 특성에 따라 j개의 범주로 분류한다면 i소지역의 총계에 대한 합성 추정량은 다음과 같이 표현된다.

$$\hat{Y}_{i.}^{S} = \sum_{j} \left(\frac{X_{ij}}{X_{.j}}\right) \hat{Y}_{.j} \qquad 2\text{-}(1)$$

여기서, X_{ij} : i소지역의 j범주의 보조 정보값

$X_{.j} = \sum_{i} X_{ij}$: 대지역에서 j범주에 대한 보조 정보값

$Y_{.j} = \sum_{i} Y_{ij}$: 대지역에서 j범주에 대한 총계

$\hat{Y}_{.j}$: 대지역에서 j범주에 대한 총계 $Y_{.j}$의 직접추정량

여기서, $\hat{Y}_{.j}$는 비추정량 형태이기 때문에 다음과 $\hat{Y}_{.j} = (\frac{y_{.j}}{x_{.j}})X_{.j}$ 로 쓸 수 있다. $y_{.j}$는 j범주에서 추출된 y의 표본총계, $x_{.j}$는 j범주에서 추출된 x의 표본총계이다.

합성추정량 $\hat{Y}_{i.}^{S}$의 정확성에 대한 측도로는 일반적으로

$$MSE(\hat{Y}_{i.}^{S}) = Var(\hat{Y}_{i.}^{S}) + [Bias(\hat{Y}_{i.}^{S})]^2 \qquad 2\text{-}(2)$$

과 같은 식으로 주어지는 합성추정량의 평균제곱오차를 고려한다.

$\hat{Y}_{i.}$가 i소지역 총계 Y_i의 직접추정량일 때, $\hat{Y}_{i.}$과 $\hat{Y}_{i.}^{S}$의 공분산 $cov(\hat{Y}_i, \hat{Y}_{i.}^{S}) = 0$ 이라는 가정하에서 $\hat{Y}_{i.}^{S}$의 MSE(mean squared error)의 근사적 불편 추정량은 다음과 같이 쓸 수 있다.

$$mse(\hat{Y}_{i.}^{S}) = (\hat{Y}_{i.}^{S} - \hat{Y}_{i.})^2 - var(\hat{Y}_{i.}) \qquad 2\text{-}(3)$$

그러나 이 추정량은 표본조사 수가 충분히 크지 않을 때에는 불안정한(unstable) 경향을 보인다.

2) 복합 추정량(Composite Estimator)

소지역 추정에서 직접추정량을 사용하게 될 경우 표본이 충분하지 않기 때문에 추정오차가 상당히 커질 수밖에 없다. 또한 대지역의 정보를 이용하여 간접적으로 추정되는 합성추정량은 편향될(biased) 가능성이 존재 한다. 따라서, 이를 보완하기 위한 방법으로 추정의 정도를 높이기 위하여 직접추정량과 합성추정량의 가중 평균을 사용하는데 이를 복합추정량이라 한다.

복합추정량의 일반적인 형태는 다음과 같다.

$$\hat{Y}_{i.}^{C} = w_i \hat{Y}_{i.} + (1 - w_i)\hat{Y}_{i.}^{S} \qquad 2\text{-}(4)$$

여기서 $\hat{Y}_{i.}$는 직접추정량이고, $\hat{Y}_{i}^{S}$는 합성추정량을 나타낸다. w_i는 0과 1사이의 가중값이다. 이 추정량에 대한 MSE는 다음과 같이 주어진다.

$$\begin{aligned} MSE(\hat{Y}_{i.}^{C}) = w_i^2 MSE(\hat{Y}_{i.}) + (1 - w_i)^2 MSE(\hat{Y}_{i.}^{S}) \\ + 2w_i(1 - w_i)E(\hat{Y}_{i.} - Y_{i.}^{*})E(\hat{Y}_{i.}^{S} - Y_{i.}^{*}) \end{aligned} \qquad 2\text{-}(5)$$

여기서, $Y_{i.}^{*}$는 소지역 i에 대한 모수 값이다. 위의 MSE가 w_i에 대하여 최소로 하는 최적 w_i는 아래와 같다.

$$w_{i(opt)}^{*} = \frac{MSE(\hat{Y}_{i.}^{S}) - E(\hat{Y}_{i.} - Y_{i.}^{*})E(\hat{Y}_{i.}^{S} - Y_{i.}^{*})}{MSE(\hat{Y}_{i.}^{S}) + MSE(\hat{Y}_{i.}) - 2E(\hat{Y}_{i.} - Y_{i.}^{*})E(\hat{Y}_{i.}^{S} - Y_{i.}^{*})} \qquad 2\text{-}(6)$$

여기서 직접추정량은 비편향 추정량이기 때문에 직접추정량의 MSE는 직접추정량의 분산과 같다. 이제, $cov(\hat{Y}_{i.}, \hat{Y}_{i.}^{S}) = 0$이라고 가정한다면 $\hat{Y}_{i.}^{c}$의 MSE를 최소로하는 최적 가중치 $w_{i(opt)}$를 근사적으로 다음과 같이 구할 수 있다.

$$w_{i(opt)} = \frac{MSE(\hat{Y}_{i.}^{S})}{MSE(\hat{Y}_{i.}^{S}) + Var(\hat{Y}_{i.})} \qquad 2\text{-}(7)$$

식 2-(7)의 최적가중치 $w_{i(opt)}$는 다음과 같이 추정할 수 있다.

$$\hat{w}_{i(opt)} = \frac{mse(\hat{Y}_{i.}^{S})}{mse(\hat{Y}_{i.}^{S}) + Var(\hat{Y}_{i.})} \qquad 2\text{-}(8)$$

따라서 식 (2.9)의 추정된 가중치를 이용하면 복합추정량은 다음과 같이 구할 수 있다.

$$\hat{Y}_{i.}^{C} = \hat{w}_{i(opt)}\hat{Y}_{i.} + (1-\hat{w}_{i(opt)})\hat{Y}_{i.}^{S} \qquad 2\text{-}(9)$$

복합추정량 $\hat{Y}_{i.}^{C}$의 정확성에 대한 측도로는 일반적으로

$$MSE(\hat{Y}_{i.}^{C}) = Var(\hat{Y}_{i.}^{C}) + [Bias(\hat{Y}_{i.}^{C})]^2 \qquad 2\text{-}(10)$$

과 같은 식으로 주어지는 복합추정량의 평균제곱오차를 고려한다. $\hat{Y}_{i.}^{C}$의 MSE(mean squared error)의 근사적 불편 추정량은 다음과 같이 쓸 수 있다.

$$mse(\hat{Y}_{i.}^{C}) = \widehat{Var}(\hat{Y}_{i.}^{C}) + [\widehat{Bias}(\hat{Y}_{i.}^{C})]^2 \qquad 2\text{-}(11)$$

여기서, $\widehat{Var}(\hat{Y}_{i.}^{C}) = \frac{n_i - 1}{n_i}\sum_{h=1}^{n_i}\left(\hat{Y}_{i.}^{C}(h) - \frac{1}{n_i}\sum_{l=1}^{n_i}\hat{Y}_{i.}^{C}(l)\right)^2$

$$\widehat{Bias}(\hat{Y}_{i.}^{C}) = (n_i - 1)\left(\frac{1}{n_i}\sum_{h=1}^{n_i}\hat{Y}_{i.}^{C}(h) - \hat{Y}_{i.}^{C}\right)$$

$$\hat{Y}_{i.}^{C}(h) = \hat{w}_{i(opt)}\hat{Y}_{i.} + (1-\hat{w}_{i(opt)})\hat{Y}_{i.}^{S}(h)$$

3. 선형 모형을 이용한 추정을 위한 이론

1) 일반선형 혼합 모형

표본자료가 일반선형혼합모형을 따른다고 가정하면 다음과 같이 일반선형 모형을 설명할 수 있다.

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{v} + \mathbf{e} \qquad 3\text{-}(1)$$

여기서 $\mathbf{y}$ 는 $n \times 1$ 표본 관측치 벡터이고, $\mathbf{X}$ 와 $\mathbf{Z}$ 는 알려진 $n \times p$와 $n \times h$ 인 완전계수 행렬이다. 그리고 $\mathbf{v}$ 와 $\mathbf{e}$ 는 서로독립으로 평균 0과 분산계수 $\delta = (\delta_1, \cdots, \delta_q)^T$에 의한 공분산행렬 $\mathbf{G}$ 와 $\mathbf{R}$ 로 나뉘어 진다. $Var(\mathbf{y})$ 는 $\mathbf{y}$ 의 분산-공분산행렬로 표시 된다.

$$Var(\mathbf{y}) = \mathbf{V} = \mathbf{V}(\delta) = \mathbf{R} + \mathbf{Z}\mathbf{G}\mathbf{Z}^T \qquad 3\text{-}(2)$$

$\mathbf{y}$ 의 선형조합은 다음과 같다.

$$\mu = \mathbf{l}^T\beta + \mathbf{m}^T\mathbf{v} \qquad 3\text{-}(3)$$

이 추정에서는 회귀계수 β와 상세벡터 $\mathbf{v}$ 를 구현, 그리고 상수 $\mathbf{l}$ 와 $\mathbf{m}$ 얻는데 관심을 갖게 된다. μ의 선형 추정량은 알려진 상수 $\mathbf{a}$ 와 $\mathbf{b}$ 의 표현식으로 표현할 수 있다.

$$\hat{\mu} = \mathbf{a}^T\mathbf{y} + \mathbf{b} \qquad 3\text{-}(4)$$

이 표현식은 μ의 불편모형식이고 각각의 기댓값과 $\hat{\mu}$의 MSE는 다음과 같이 주어진다.

$$E(\hat{\mu}) = E(\mu) \qquad 3\text{-}(5)$$

$$MSE(\hat{\mu}) = E(\hat{\mu} - \mu)^2 \qquad 3\text{-}(6)$$

만약 $\hat{\mu}$가 μ의 불편이면, 오차 $\hat{\mu} - \mu$의 분산으로 정리할 수 있다.

$$MSE(\hat{\mu}) = Var(\hat{\mu} - \mu) \qquad 3\text{-}(7)$$

이제, 불편선형 추정량 $\hat{\mu}$의 MSE를 최소로 하는 BLUP(Best Linear Unbiased Prediction) 추정량을 찾는데 관심을 두게 된다.

2) BLUP 추정량

알려진 δ에 의해, μ의 BLUP 추정량은 다음과 같이 주어진다.

$$\tilde{\mu}^H = t(\delta, \mathbf{y}) = \mathbf{l}^T\tilde{\beta} + \mathbf{m}^T\tilde{\mathbf{v}} = \mathbf{l}^T\tilde{\beta} + \mathbf{m}^T\mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\beta}) \qquad 3\text{-}(8)$$

β의 최적선형 비편향 추정량 $\tilde{\beta}$는 다음과 같이 주어진다.

$$\tilde{\beta} = \tilde{\beta}(\delta) = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \qquad 3\text{-}(9)$$

그리고, $\tilde{\mathbf{v}}$는 다음과 같다.

$$\tilde{\mathbf{v}} = \tilde{\mathbf{v}}(\delta) = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\beta}) \qquad 3\text{-}(10)$$

μ의 최적예측(BP) 추정량은 조건부기대값 $E(\mu|\mathbf{y})$에 의해 다음과 같이 주어진다.

$$E[d(\mathbf{y})-\mu]^2 \geqq E[E(\mu|\mathbf{y})-\mu]^2 \qquad 3\text{-}(11)$$

μ의 다른 추정량 $d(\mathbf{y})$는 반드시 선형이거나 불편성을 가지지는 않는다. 이 결과에 의해 $d(\mathbf{y})=E(\mu|\mathbf{y})$이라는 가정하에서 다음식을 쓸 수 있다.

$$\begin{aligned} E[(d(\mathbf{y})-\mu)^2|\mathbf{y}] &>= E[(d(\mathbf{y})-E(\mu|\mathbf{y})+E(\mu|\mathbf{y})-\mu)^2|\mathbf{y}] \\ &= [d(\mathbf{y})-E(\mu|\mathbf{y})]^2+E[(E(\mu|\mathbf{y})-\mu)^2|\mathbf{y}] \\ &\geqq E[(E(\mu|\mathbf{y})-\mu)^2|y] \end{aligned} \qquad 3\text{-}(12)$$

정규성 하에서, BP 추정량의 $E(\mu|\mathbf{y})$는 $\boldsymbol{\beta}$를 대신하여 식3-(9)의 $\tilde{\boldsymbol{\beta}}$를 갖는 BLUP 추정량으로 바뀐다. 그것은 알려지지 않은 $\boldsymbol{\beta}$에 의한다. 특히,

$$E(\mathbf{m}^T\mathbf{v}|\mathbf{y})=\mathbf{m}^T\mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \qquad 3\text{-}(13)$$

이고, 이 추정량은 $E(\mathbf{m}^T\mathbf{v}|\mathbf{y})$의 정규성 가정이 없는 최적선형예측(BLP) 추정량이 된다.

$\boldsymbol{\beta}$에 의한 추정량을 원하지 않을 경우, $\mathbf{y}$를 평균 0을 따르는 오차 대조 $\mathbf{A}^T\mathbf{y}$로 바꿀수 있다. 여기서 $\mathbf{A}^T\mathbf{X}=0$이고 $\mathbf{A}$는 $n\times p$인 모형 행렬 $\mathbf{X}$에 직교하는 $n\times(n-p)$인 완전계수 행렬이다. 바꾼 자료에서 최적 예측 $E(\mathbf{m}^T\mathbf{v}|\mathbf{A}^T\mathbf{y})$은 사실상 BLUP 추정량 $\mathbf{m}^T\tilde{\mathbf{v}}$으로 변형되게 된다. 이러한 결과는 정규성 가정에서, 선형성이나 불편성없이 BLUP의 정당성을 나타낸다.

여기서는, 단순 회귀 조합 $\mu=\mathbf{l}^T\boldsymbol{\beta}+\mathbf{m}^T\mathbf{v}$의 BLUP 추정을 고려하지만, 쉽게 $r(\geqq 2)$ 선형 조합 $\boldsymbol{\mu}=\mathbf{L}\boldsymbol{\beta}+\mathbf{M}\mathbf{v}$, $\boldsymbol{\mu}=(\mu_1,\cdots,\mu_r)^T$의 추정으로 확장되어 진다. $\boldsymbol{\mu}$의 BLUP 추정량은 다음과 같다.

$$\mathbf{t}(\delta,\mathbf{y}) = \mathbf{L}\tilde{\beta} + \mathbf{M}\tilde{\mathbf{v}} = \mathbf{L}\tilde{\beta} + \mathbf{MGZ}^{T}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\beta}) \quad 3\text{-}(14)$$

추정량 $\mathbf{t}(\delta,\mathbf{y})$ 는 μ 의 다른 선형 비편향 추정량 $\mathbf{t}^{*}(\mathbf{y})$ 에서 최적이다. 여기서 $E(\mathbf{t}^{*}-\mu)(\mathbf{t}^{*}-\mu)^{T} - E(\mathbf{t}-\mu)(\mathbf{t}-\mu)^{T}$는 양의 부정준부호 행렬(psd)이고, $E(\mathbf{t}-\mu)(\mathbf{t}-\mu)^{T}$는 $\mathbf{t}-\mu$ 의 공분산 행렬이 된다.

(1) BLUP의 MSE

BLUP 추정량 $\mathbf{t}(\delta,\mathbf{y})$ 는 다음과 같이 표현되기도 한다.

$$\mathbf{t}(\delta,\mathbf{y}) = \mathbf{t}^{*}(\delta,\beta,\mathbf{y}) + \mathbf{d}^{T}(\tilde{\beta} - \beta) \quad 3\text{-}(15)$$

여기서 $\mathbf{t}^{*}(\delta,\beta,\mathbf{y})$ 는 β 가 알려졌을 때의 BLUP 추정량이다.

$$\mathbf{t}^{*}(\delta,\beta,\mathbf{y}) = \mathbf{l}^{T}\beta + \mathbf{b}^{T}(\mathbf{y} - \mathbf{X}\beta) \quad 3\text{-}(16)$$

$\mathbf{b}^{T}$와 $\mathbf{d}^{T}$는 다음과 같다.

$$\mathbf{b}^{T} = \mathbf{m}^{T}\mathbf{GZ}^{T}\mathbf{V}^{-1},\ \mathbf{d}^{T} = \mathbf{l}^{T} - \mathbf{b}^{T}\mathbf{X} \quad 3\text{-}(17)$$

$\mathbf{t}^{*}(\delta,\beta,\mathbf{y}) - \mu$ 와 $\mathbf{d}^{T}(\tilde{\beta} - \beta)$ 는 서로 연관되어 있지 않으며 다음과 같이 쓸 수 있다.

$$E[(\mathbf{b}^{T}(\mathbf{Zv} + \mathbf{e}) - \mathbf{m}^{T}\mathbf{v})(\mathbf{v}^{T}\mathbf{Z}^{T} + \mathbf{e}^{T})\mathbf{V}^{-1}] = \mathbf{0} \quad 3\text{-}(18)$$

따라서, $MSE[\mathbf{t}(\delta,\mathbf{y})]$ 는

$$MSE[\mathbf{t}(\delta,\mathbf{y})] = MSE[\mathbf{t}^{*}(\delta,\beta,\mathbf{y})] + Var[\mathbf{d}^{T}(\tilde{\beta} - \beta)] = g_1(\delta) + g_2(\delta) \quad 3\text{-}(19)$$

가 된다. 여기서, $g_1(\delta)$와 $g_2(\delta)$는

$$g_1(\delta) = Var[\mathbf{t}^*(\delta,\beta,\mathbf{y}) - \mu] = \mathbf{m}^T(\mathbf{G} - \mathbf{GZ}^T\mathbf{V}^{-1}\mathbf{ZG})\mathbf{m} \quad 3\text{-}(20)$$

$$g_2(\delta) = \mathbf{d}^T(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{d} \quad 3\text{-}(21)$$

가 된다. 식 3-(19)에서, 두 번째 항의 $g_2(\delta)$는 추정량 $\tilde{\beta}$ 에서의 변이성을 설명한다.

3) EBLUP 추정량

식 3-(8)에서 BLUP 추정량 $\mathbf{t}(\delta,\mathbf{y})$ 는 실제 응용에서 알려지지 않는 분산 계수 δ 에 의존한다. δ를 대신해 추정량 $\hat{\delta} = \hat{\delta}(\mathbf{y})$ 으로 바꾸면 EBLUP 추정량으로 나타내어 지는 2단계 추정량 $\hat{\mu}^H = t(\hat{\delta},\mathbf{y})$를 얻을 수 있다. $\mathbf{t}(\delta,\mathbf{y})$ 와 $\mathbf{t}(\hat{\delta},\mathbf{y})$ 처럼 $\mathbf{t}(\delta)$ 와 $\mathbf{t}(\hat{\delta})$ 로 쓸 수 있다.

2단계 추정량 $\mathbf{t}(\hat{\delta})$ 는 μ에 비편향이라는 성질이 남아 $E[\mathbf{t}(\hat{\delta}) - \mu] = 0$ 가 된다. 이것은 다음 3가지 성질을 가지게 된다. (1) $E[\mathbf{t}(\hat{\delta})]$가 유한하다, (2) $\hat{\delta}$ 가 δ의 전이불변이라는 이라는 것이다. 즉, 모든 $\mathbf{y}$ 와 $\mathbf{b}$ 에 대해 $\hat{\delta}(-\mathbf{y}) = \hat{\delta}(\mathbf{y})$ 이고, $\hat{\delta}(\mathbf{y} - \mathbf{Xb}) = \hat{\delta}(\mathbf{y})$ 이라는 것이다. (3) $\mathbf{v}$ 와 $\mathbf{e}$ 의 분포는 정규분포가 아니라고 하더라도 0 주변에서 서로 대칭이다.

식 3-(1)의 일반선형혼합모형 특별한 형태인 분산분석 방식에 의한 ANOVA 모형은 다음과 같이 주어진다.

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}_1\mathbf{v}_1 + \cdots \mathbf{Z}_r\mathbf{v}_r + \mathbf{e} \quad 3\text{-}(22)$$

여기서 $\mathbf{v}_1, \cdots, \mathbf{v}_r$ 과 $\mathbf{e}$ 는 독립이고 각각 평균 0과 공분산 행렬 $\sigma_1^2\mathbf{I}_{h_1}, \cdots, \sigma_r^2\mathbf{I}_{h_r}$와 $\sigma_e^2\mathbf{I}_n$ 의 분포를 갖는다. $\sigma_i^2 \geqq 0 (i = 1, \cdots, r)$과 $\sigma_0^2 = \sigma_e^2 > 0$ 에 의한 계수

$\boldsymbol{\delta} = (\sigma_0^2, \cdots, \sigma_r^2)^T$는 분산성분이다. 여기에서 $\mathbf{G}$ 는 블록 $\sigma_i^2 \mathbf{I}_{h_i}$의 블록대각행렬이고, $\mathbf{R} = \sigma_e^2 \mathbf{I}_n$ 이고 $\mathbf{V} = \sigma_e^2 \mathbf{I}_n + \sum \sigma_i^2 \mathbf{Z}_i \mathbf{Z}_i{}^T$는 선형 구조에서 공분산행렬의 특별한 형태가 된다. : 대칭행렬 $\mathbf{H}_i$ 에 의해 $\mathbf{V} = \sum \delta_i \mathbf{H}_i$ 가 된다.

(1) EBLUP의 MSE

EBLUP 추정량 $\mathbf{t}(\hat{\delta})$ 에서의 오차는 다음과 같이 분해된다.

$$t(\hat{\delta}) - \mu = [t(\delta) - \mu] + [t(\hat{\delta}) - t(\delta)] \qquad 3\text{-}(23)$$

따라서, $MSE[t(\hat{\delta})]$는 다음과 같다.

$$MSE[t(\hat{\delta})] = MSE[t(\delta)] + E[t(\hat{\delta}) - t(\delta)]^2 + 2E[t(\delta) - \mu][t(\hat{\delta}) - t(\delta)] \qquad 3\text{-}(24)$$

랜덤 효과 $\mathbf{v}$ 와 $\mathbf{e}$ 의 정규성 하에서, 식 3-(24)에서의 교호항은 0이고 $\hat{\delta}$ 는 전이불변이다. 따라서, $MSE[t(\hat{\delta})]$는 다음과 같다.

$$MSE[t(\hat{\delta})] = MSE[t(\delta)] + E[t(\hat{\delta}) - t(\delta)]^2 \qquad 3\text{-}(25)$$

정규성하에서, 식 3-(25)에서 EBLUP추정량의 MSE는 BLUP 추정량 $t(\delta)$의 MSE보다 항상 크게 된다. $MSE[t(\delta)]$에 의한 $MSE[t(\hat{\delta})]$의 근사치의 일반적인 실행은 상당한 과소추정이 따르고, 특히, 이 경우에 상당한 정도의 δ 의해 $t(\delta)$가 변하고, $\hat{\delta}$ 의 변이성이 발생하게 된다.

식 3-(25)의 마지막 항은 1원 모형 $y_{ij} = \mu + v_i + e_{ij}, i = 1, \cdots, m, j = 1, \cdots, \overline{n}$ 과 같은 특별한 경우를 제외하고 일반적으로 수정하기 어렵다. 따라서, 이 식에서

경험적 근사치를 얻을 필요성이 있고, 다음식을 얻을 수 있다.

$$t(\hat{\delta}) - \mu \approx \mathbf{d}(\delta)^{T}(\hat{\delta} - \delta) \qquad 3\text{-}(26)$$

여기서 $\mathbf{d}(\delta) = \partial t(\delta)/\partial\delta$ 이고, $\hat{\delta} - \delta$ 의 고차항 제곱을 포함하는 항은 $\mathbf{d}(\delta)^{T}(\hat{\delta} - \delta)$ 에 관계되는 낮은 차수가 됨을 가정한다. 또한, 정규성 가정하에서 $\mathbf{d}(\delta)$ 는 다음과 같이 쓸 수 있다.

$$\mathbf{d}(\delta) \approx \partial t^{*}(\delta, \beta)/\partial\delta = (\partial \mathbf{b}^{T}/\partial\delta)(\mathbf{y} - \mathbf{X}\beta) = \mathbf{d}^{*}(\delta) \qquad 3\text{-}(27)$$

δ 에 대하여 $\tilde{\beta} - \beta$ 의 도함수를 포함하는 식은 낮은 차수를 가지고, 여기서 $t^{*}(\delta, \beta)$는 식 3-(16)에서 구할수 있고, 따라서, 다음과 같이 쓸수 있다.

$$E[\mathbf{d}(\delta)^{T}(\hat{\delta} - \delta)]^{2} \approx E[\mathbf{d}^{*}(\delta)^{T}(\hat{\delta} - \delta)]^{2} \qquad 3\text{-}(28)$$

또한, $E[\mathbf{d}^{*}(\delta)^{T}(\hat{\delta} - \delta)]^{2}$ 는

$$\begin{aligned} E[\mathbf{d}^{*}(\delta)^{T}(\hat{\delta} - \delta)]^{2} &\approx tr[E(\mathbf{d}^{*}(\delta)\mathbf{d}^{*}(\delta)^{T})\overline{\mathbf{V}}(\hat{\delta})] \\ &= tr[(\partial \mathbf{b}^{T}/\partial\delta)\mathbf{V}(\partial \mathbf{b}^{T}/\partial\delta)^{T}\overline{\mathbf{V}}(\hat{\delta})] =: g_{3}(\delta) \end{aligned} \qquad 3\text{-}(29)$$

가 된다. 여기서 낮은 차수가 배제된 식 $\overline{\mathbf{V}}(\hat{\delta})$ 는 $\hat{\delta}$ 의 점근 공분산행렬 이고 A=:B는 B가 A와 같다는 것을 의미한다. 이에 따라 식 3-(26), 3-(28), 3-(29)에 의해 다음과 같이 쓸 수 있다.

$$E[\mathbf{d}(\delta)^{T}(\hat{\delta} - \delta)]^{2} \approx g_{3}(\delta) \qquad 3\text{-}(30)$$

식 3-(20), 3-(21)과 식 3-(30)을 조합하면, $t(\hat{\delta})$의 MSE의 2차 근사치를 얻을 수 있다.

$$MSE[t(\hat{\delta})] \approx g_1(\delta) + g_2(\delta) + g_3(\delta) \qquad 3\text{-}(31)$$

$g_2(\delta)$와 $g_3(\delta)$ 항은 β와 δ에 기인하고 첫째 항 $g_1(\delta)$ 보다 낮은 차수를 갖는다.

(2) EBLUP의 MSE 추정량

실제 응용에서는, $t(\hat{\delta})$ 변이성 계산에 의한 $MSE[t(\hat{\delta})]$ 추정량이 필요하게 된다. $MSE[t(\hat{\delta})]$가 $MSE[t(\delta)]$의 근사치에 접근하고, δ 대신에 $\hat{\delta}$을 사용하면, MSE의 추정량은 다음과 같이 주어진다.

$$mse_N[t(\hat{\delta})] = g_1(\hat{\delta}) + g_2(\hat{\delta}) \qquad 3\text{-}(32)$$

다른 MSE의 추정량은 MSE 근사치 식 3-(31)에 δ 대신에 $\hat{\delta}$을 사용하면 다음 식을 얻을 수 있다.

$$mse_1[t(\hat{\delta})] = g_1(\hat{\delta}) + g_2(\hat{\delta}) + g_3(\hat{\delta}) \qquad 3\text{-}(33)$$

근사치에 의한 차수에 따라 $Eg_2(\hat{\delta}) \approx g_2(\delta)$와 $Eg_3(\hat{\delta}) \approx g_3(\delta)$가 되지만, $g_1(\hat{\delta})$는 $g_1(\delta)$의 최적 추정량이 되지 않는데 그 이유는 일반적으로 $g_2(\delta)$와 $g_3(\delta)$에 따라 같은 차수의 편향이 있기 때문이다.

$g_1(\hat{\delta})$의 편향을 구하기 위해, δ에 의한 $g_1(\hat{\delta})$를 확장시킬 수 있다.

$$g_1(\hat{\boldsymbol{\delta}}) = g_1(\boldsymbol{\delta}) + (\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})^T \nabla g_1(\boldsymbol{\delta}) + \frac{1}{2}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})^T \nabla^2 g_1(\boldsymbol{\delta})(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}) \quad 3\text{-}(34)$$
$$= g_1(\boldsymbol{\delta}) + \triangle_1 + \triangle_2$$

여기서 $\nabla g_1(\boldsymbol{\delta})$는 $\boldsymbol{\delta}$에 관계된 $g_1(\boldsymbol{\delta})$의 일차도함수의 벡터가 되고, $\nabla^2 g_1(\boldsymbol{\delta})$는 $\boldsymbol{\delta}$에 관계된 $g_1(\boldsymbol{\delta})$의 이차도함수 행렬이 된다. 만약 $\hat{\boldsymbol{\delta}}$가 $\boldsymbol{\delta}$의 비편향이라면, $E(\triangle_1) = 0$이다. 일반적으로, $E(\triangle_1) \approx \mathbf{b}_{\hat{\delta}}^{\mathbf{T}}(\boldsymbol{\delta}) \nabla\, g_1(\boldsymbol{\delta})$는 $E(\triangle_2)$보다 낮은 차수가 된다. 그러면,

$$Eg_1(\hat{\boldsymbol{\delta}}) \approx g_1(\boldsymbol{\delta}) + \frac{1}{2} tr[\nabla^2 g_1(\boldsymbol{\delta}) \overline{\mathbf{V}}(\hat{\boldsymbol{\delta}})] \quad 3\text{-}(35)$$

여기서 $\mathbf{b}_{\hat{\delta}}(\boldsymbol{\delta})$는 $E(\hat{\boldsymbol{\delta}}) - \boldsymbol{\delta}$ 편향의 근사치가 된다. 만약 공분산행렬 $\mathbf{V}$가 선형 구조를 가지고 있다면, 식 3-(35)은 다음과 같이 쓸 수 있다.

$$Eg_1(\hat{\boldsymbol{\delta}}) \approx g_1(\boldsymbol{\delta}) - g_3(\boldsymbol{\delta}) \quad 3\text{-}(36)$$

이제 식 3-(32), 3-(33)와 식 3-(36)으로부터, $mse_N[t(\hat{\boldsymbol{\delta}})]$와 $mse_1[t(\hat{\boldsymbol{\delta}})]$의 편향은 다음과 같다.

$$B_N \approx -2g_3(\boldsymbol{\delta}),\ B_1 \approx -g_3(\boldsymbol{\delta}) \quad 3\text{-}(37)$$

근사치의 차수를 따르는 $MSE[t(\hat{\boldsymbol{\delta}})]$의 최적 추정량은 다음과 같이 주어진다.

$$mse[t(\hat{\boldsymbol{\delta}})] \approx g_1(\hat{\boldsymbol{\delta}}) + g_2(\hat{\boldsymbol{\delta}}) + 2g_3(\hat{\boldsymbol{\delta}}) \quad 3\text{-}(38)$$

식 3-(37)로부터 $E[g_1(\hat{\boldsymbol{\delta}}) + g_3(\hat{\boldsymbol{\delta}})] \approx g_1(\boldsymbol{\delta})$ 임을 알 수 있고, 결과적으로

$$Emse[t(\hat{\boldsymbol{\delta}})] \approx MSE[t(\hat{\boldsymbol{\delta}})] \qquad 3\text{-}(39)$$

가 된다.

(3) 블록 대각공분산 구조

일반선형 혼합 모형 식 3-(1)의 특별한 형태로 많은 소지역 모형을 포함하는 일반선형 혼합 모형을 설명할 수 있다. 이 모형은 다음의 블록 대각 공분산 구조를 가진다.

$$\mathbf{y} = col_{1 \leq i \leq m}(\mathbf{y}_i) = (\mathbf{y}_1^T, \cdots, \mathbf{y}_m^T)^T,\ \mathbf{X} = col_{1 \leq i \leq m}(\mathbf{X}_i), \qquad 3\text{-}(40)$$
$$\mathbf{Z} = diag_{1 \leq i \leq m}(\mathbf{Z}_i),\ \mathbf{v} = col_{1 \leq i \leq m}(\mathbf{v}_i),\ \mathbf{e} = col_{1 \leq i \leq m}(\mathbf{e}_i)$$

여기서 m은 소지역의 수이고, $\mathbf{X}_i$는 $n_i \times p$이고, $\mathbf{Z}_i$는 $n_i \times h_i$이며, $\mathbf{y}_i$는 $n_i \times 1$ 벡터이다($\sum n_i = n$, $\sum h_i = h$).

$$\mathbf{R} = diag_{1 \leq i \leq m}(\mathbf{R}_i),\ \mathbf{G} = diag_{1 \leq i \leq m}(\mathbf{G}_i) \qquad 3\text{-}(41)$$

$\mathbf{V}$는 블록 대각 구조를 가지고 $\mathbf{V}$는

$$\mathbf{V} = diag_{1 \leq i \leq m}(\mathbf{V}_i) \qquad 3\text{-}(42)$$

이고, $\mathbf{V}_i$는 다음과 같다.

$$\mathbf{V}_i = \mathbf{R}_i + \mathbf{Z}_i \mathbf{G}_i \mathbf{Z}_i^T \qquad 3\text{-}(43)$$

따라서, m 단위모형으로 분해되어 지고 다음과 같이 쓸 수 있다.

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{v}_i + \mathbf{e}_i, i = 1, \cdots, m \qquad 3\text{-}(44)$$

이제 다음의 선형 조합 추정식에 관심을 갖게 된다.

$$\boldsymbol{\mu}_i = \mathbf{l}_i^T\boldsymbol{\beta} + \mathbf{m}_i^T\mathbf{v}_i, i = 1, \cdots, m \qquad 3\text{-}(45)$$

식 3-(8)에 따라 $\boldsymbol{\mu}_i$의 BLUP 추정량은 다음과 같이 바꿀수 있고

$$\tilde{\boldsymbol{\mu}}_i^H = t_i(\boldsymbol{\delta}, \mathbf{y})_i = \mathbf{l}_i^T\tilde{\boldsymbol{\beta}} + \mathbf{m}_i^T\tilde{\mathbf{v}}_i, i = 1, \cdots, m \qquad 3\text{-}(46)$$

여기서, $\tilde{\mathbf{v}}_i$는

$$\tilde{\mathbf{v}}_i = \mathbf{G}_i\mathbf{Z}_i^T\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\tilde{\boldsymbol{\beta}}) \qquad 3\text{-}(47)$$

이다. 그리고 $\tilde{\boldsymbol{\beta}}$ 는

$$\tilde{\boldsymbol{\beta}} = (\sum_i \mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1}(\sum_i \mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{y}_i) \qquad 3\text{-}(48)$$

이다. BLUP의 MSE 추정량은 식 3-(19)에 의해 다음과 같이 바꿀 수 있다.

$$MSE(\tilde{\boldsymbol{\mu}}_i^H) = g_{1i}(\boldsymbol{\delta}) + g_{2i}(\boldsymbol{\delta}) \qquad 3\text{-}(49)$$

여기서 $g_{1i}(\boldsymbol{\delta})$와 $g_{2i}(\boldsymbol{\delta})$는

$$g_{1i}(\boldsymbol{\delta}) = \mathbf{m}_i^T(\mathbf{G}_i - \mathbf{G}_i\mathbf{Z}_i^T\mathbf{v}_i^{-1}\mathbf{Z}_i\mathbf{G}_i)\mathbf{m}_i \quad 3\text{-}(50)$$

$$g_{2i}(\boldsymbol{\delta}) = \mathbf{d}_i^T(\sum_i \mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1}\mathbf{d}_i \quad 3\text{-}(51)$$

이고, $\mathbf{d}_i^T$와 $\mathbf{b}_i^T$는 다음과 같다.

$$\mathbf{d}_i^T = \mathbf{l}_i^T - \mathbf{b}_i^T\mathbf{X}_i \quad 3\text{-}(52)$$

$$\mathbf{b}_i^T = \mathbf{m}_i^T\mathbf{G}_i\mathbf{Z}_i^T\mathbf{V}_i^{-1} \quad 3\text{-}(53)$$

$\boldsymbol{\delta}$를 대신하여 $\hat{\boldsymbol{\delta}}$ 추정량으로 바꾸면, EBLUP 추정량을 얻을 수 있다.

$$\hat{\boldsymbol{\mu}}_i^H = t_i(\hat{\boldsymbol{\delta}}, \mathbf{y_i}) = \mathbf{l}_i^T\hat{\boldsymbol{\beta}} + \mathbf{m}_i^T\hat{\mathbf{v}}_i, i = 1, \cdots, m \quad 3\text{-}(54)$$

2차 MSE 근사치 식 3-(31)는 다음과 같이 바꿀 수 있다.

$$MSE(\hat{\boldsymbol{\mu}}_i^H) \approx g_{1i}(\boldsymbol{\delta}) + g_{2i}(\boldsymbol{\delta}) + g_{3i}(\boldsymbol{\delta}) \quad 3\text{-}(55)$$

여기서 $g_{3i}(\boldsymbol{\delta})$는

$$g_{3i}(\boldsymbol{\delta}) = tr[(\partial\mathbf{b}_i^T/\partial\boldsymbol{\delta})\mathbf{V}_i(\partial\mathbf{b}_i^T/\partial\boldsymbol{\delta})^T\overline{\mathbf{V}}(\hat{\boldsymbol{\delta}})] \quad 3\text{-}(56)$$

이고, MSE 추정량은 식 3-(38)에 의해 다음과 같이 바뀐다.

$$mse(\hat{\boldsymbol{\mu}}_i^H) \approx g_{1i}(\hat{\boldsymbol{\delta}}) + g_{2i}(\hat{\boldsymbol{\delta}}) + 2g_{3i}(\hat{\boldsymbol{\delta}}) \quad 3\text{-}(57)$$

4. 회귀모형에 의한 추정량의 적용

1) 내포오차 선형 회귀모형

내포오차 선형 회귀모형은 소지역의 관심변수를 추정하기 위한 모형이다. 그 모집단 모형은 다음과 같이 주어진다.

$$y_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + v_i + e_{ij},\ j = 1, \cdots, N_i,\ \ i = 1, \cdots, m \qquad 4\text{-}(1)$$

여기서, y_{ij} : i 번째 소지역의 j번째의 모집단 단위(관심변수)

$\boldsymbol{x}'_{ij}$: y_{ij}와 연관된 $x_{ij1} = 1$ 이고 $p \times 1$인 보조변수 벡터

$\boldsymbol{x}'_{ij} = (x_{ij1}, \cdots, x_{ijp})'$

$\boldsymbol{\beta}$: $p \times 1$인 회귀계수 벡터 $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_{p-1})'$

N_i : i 번째 소지역의 모집단 단위 개수

v_i는 i의 소지역 효과를 나타내는 오차이고, e_{ij}는 소지역 i 내의 j의 표본 추출에 의한 오차를 나타낸다. v_i와 e_{ij}는 서로 독립이고 동일한 분포를 따르는 확률변수로서 다음과 같이 정의된다.

$$v_i \sim N(0, \sigma_v^2), \qquad e_{ij} \sim N(0, \sigma_e^2) \qquad 4\text{-}(2)$$

i 번째 소지역의 평균 $\overline{Y}_i$는 다음 식으로 볼 수 있다.

$$\theta_i = \overline{\boldsymbol{X}}_i'\boldsymbol{\beta} + v_i \qquad 4\text{-}(3)$$

N_i가 크다고 가정하면, $\overline{\boldsymbol{X}_i}'$ 는 i번째 지역의 알려진 모집단 평균 벡터가 된다. 즉, $\overline{X_i} = (x_{i1} + \cdots + x_{iN_i})/N_i$가 된다.

표본을 소지역간에 독립적으로 추출한다는 가정하에서 표본설계를 하고, 표본 데이터 y_{ij}, x_{ij}가 모집단 모형을 따른다고 가정하면, 다음과 같은 식을 얻을 수 있다.

$$y_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + v_i + e_{ij},\ j = 1, \cdots, n_i,\ i = 1, \cdots, m \qquad 4\text{-}(4)$$

여기서, n_i : i번재 소지역의 표본 크기,

n : 전체 표본 크기 $n = \sum_{i=1}^{m} n_i$

이 모형은 표본 설계를 무시되거나, 선택편의가 없음을 내포한다.

식 4-(4)에 의한 소지역 평균 θ_i의 모형기반 추정량은 조사 가중치 $\tilde{w}_{ij}$에 의존하지 않고, y_{ij}를 적용한다. 모형이 지역내에서 자체 가중치가 있을 때 즉, 지역내에서 단순무작위추출을 하여 가중치가 $\tilde{w}_{ij} = \tilde{w}_i$가 될 때는 제외하면, 표본크기 n_i 증가에 의한 설계일치성은 배제되어 진다. 한편, 직접설계기반 추정량은 설계일치성을 가지지만, 인접지역으로부터 영향이 없다면 설계일치성을 가지지 못하게 되며 설계기반이 아닌 모형기반 추정량이 된다. 직접설계기반 추정량의 소지역평균 θ_i는 비율의 추정량에 의해 다음과 같이 주어진다.

$$\bar{y}_{iw} = \frac{\sum_{j=1}^{n_i} \tilde{w}_{ij} y}{\sum_{j=1}^{n_i} \tilde{w}_{ij}} = \sum_{j=1}^{n_i} w_{ij} y_{ij_{ij}} \qquad 4\text{-}(5)$$

여기서, $w_{ij} = \tilde{w}_{ij} / \sum_{j=1}^{n_i} \tilde{w}_{ij} = \tilde{w}_{ij} / \tilde{w}_{i.}$, $\sum_{j=1}^{n_i} w_{ij} = 1$

4-(4)식에 4-(5)식을 적용하면, 다음의 조사가중치지역별 모형을 얻을 수 있다.

$$y_{iw} = \bar{\boldsymbol{x}}'_{iw}\boldsymbol{\beta} + v_i + \bar{e}_{iw},\ i = 1, \cdots, m \qquad 4\text{-}(6)$$

여기서, $\bar{e}_{iw} = \sum_{j=1}^{n_i} w_{ij} e_{ij},\ E(\bar{e}_{iw}) = 0\ ,\ var(\bar{e}_{iw}) = \sigma_e^2 \sum_{j=1}^{n_i} w_{ij}^2 \equiv \sigma_e^2 \delta_i^2,$

$$\bar{x}_{iw} = \sum_{j=1}^{n_i} w_{ij} x_{ij}$$

주의할 것은, 회귀계수 $\boldsymbol{\beta}$와 σ_v^2와 σ_e^2 는 두 모형 4-(4)와 4-(6)에서는 알려지지 않는다는 것이다.

2) EBLUP 추정량

(1) EBLUP 추정량

σ_v^2 와 σ_e^2가 알려져 있을 경우, 4-(4)식의 내포호차선형회귀모형에 기반한 소지역 평균 $\theta_i = \overline{\boldsymbol{X}}_i'\boldsymbol{\beta} + v_i$의 BLUP(Best Linear Unbiased Prediction) 추정량은 다음과 같이 주어진다.

$$\hat{\theta}_i = \gamma_i \bar{y}_i + (\overline{\boldsymbol{X}}_i - \gamma_i \overline{\boldsymbol{x}}_i)'\tilde{\boldsymbol{\beta}} \qquad 4\text{-}(7)$$

여기서, $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_{ij},\ \bar{x}_i = \sum_{j-1}^{n_i} x_{ij}/n_i,\ \gamma_i = \sigma_v^2/(\sigma_v^2 + \sigma_e^2/n_i)$

또한, $\tilde{\boldsymbol{\beta}}$ 는 다음과 같다.

$$\tilde{\boldsymbol{\beta}} = (\sum_{i=1}^{m} \boldsymbol{x}'_i \boldsymbol{V}_i^{-1} \boldsymbol{x}_i)^{-1} (\sum_{i=1}^{m} \boldsymbol{x}'_i \boldsymbol{V}_i^{-1} \boldsymbol{y}_i) \equiv \tilde{\boldsymbol{\beta}}\,(\sigma_e^2, \sigma_v^2) \qquad 4\text{-}(8)$$

여기서, $\boldsymbol{x}'_i = (x_{i1}, \cdots, x_{1n_i}), \boldsymbol{y}_i = (y_{i1}, \cdots, y_{1n_i})'$

$\boldsymbol{V}_i$는 다음과 같이 주어진다.

$$V_i = \sigma_e^2 I_{n_i} + \sigma_v^2 \mathbf{1}_n, \mathbf{1}'_n, \qquad 4\text{-}(9)$$

여기서, $\mathbf{1}_n$: 계수 n_i인 단위 행렬

I_n : 각 요소가 1인 계수가 n_i인 제곱행렬

$\tilde{\theta}_i$와 $\tilde{\beta}$는 σ_v^2 과 σ_e^2에 의존한다. σ_v^2 과 σ_e^2의 적합 추정량 $\hat{\sigma}_v^2$와 $\hat{\sigma}_e^2$을 사용하면. $\hat{\sigma}_v^2 = \max(\tilde{\sigma}_v^2, 0)$로 주어지고, $\tilde{\sigma}_v^2$는 다음과 같다.

$$\tilde{\sigma}_v^2 = \frac{1}{n_*}\left\{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\hat{u}_{ij}^{\,2} - (n-p)\hat{\sigma}_e^2\right\} \qquad 4\text{-}(10)$$

여기서, $n_* = n - tr\left\{(X'X)^{-1}\sum_{i=1}^{m} n_i^2 \overline{x}_i \overline{x}'_i\right\}$

$X' = (x'_1, \cdots, x'_m)$

$\hat{u}_{ij}$: $x_{ij1}, \cdots, x_{ijp}$상에서의 y_{ij}의 최소제곱 회귀의 잔차

또한, $\hat{\sigma}_e^2$는 다음과 같다.

$$\hat{\sigma}_e^2 = \nu_1^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\hat{\epsilon}_{ij}^2 \ , \qquad 4\text{-}(11)$$

여기서, $\nu_1 = n - m - p_1$, p_1 : 편차가 0인 아닌 x의 개수

$\hat{\epsilon}_{ij}$: $x_{ij1} - \overline{x}_{i.1}, \cdots x_{ijp} - \overline{x}_{i.p}$상에서의 $y_{ij} - \overline{y}_i$의 최소제곱 회귀의 잔차

식 4-(7)와 4-(8)에서 σ_v^2 과 σ_e^2 대신에 추정량 $\hat{\sigma}_v^2$ 와 $\hat{\sigma}_e^2$ 을 대입하면, 소지역 평균 θ_i의 EBLUP(Empirical Best Linear Unbiased Prediction) 추정량은 다음과 같다.

$$\hat{\theta}_i = \hat{\gamma}_i \overline{y}_i + (\overline{X}_i - \hat{\gamma}_i \overline{x}_i)'\hat{\beta}$$

여기서, $\hat{\gamma}_i = \hat{\sigma}_v^2/(\hat{\sigma}_v^2 + \hat{\sigma}_e^2/n_i)$, $\hat{\beta} = \tilde{\beta}(\hat{\sigma}_v^2, \hat{\sigma}_e^2)$

EBLUP 추정량 θ_i의 모형 MSE(mean squared error)는 다음식으로 근사한다.

$$MSE(\theta_i) \approx g_{1i}(\sigma_v^2, \sigma_e^2) + g_{2i}(\sigma_v^2, \sigma_e^2) + g_{3i}(\sigma_v^2, \sigma_e^2) \quad 4\text{-}(12)$$

여기서, $g_{1i}(\sigma_e^2, \sigma_v^2) = (1-\gamma_i)\sigma_e^2$

$$g_{2i}(\sigma_e^2, \sigma_v^2) = (\overline{\boldsymbol{X}}_{i-}\gamma_i\overline{\boldsymbol{x}}_i)'(\sum_{i=1}^{m} \boldsymbol{x}'_i \boldsymbol{V}_{i^{-1}}\boldsymbol{x}_i)^{-1}(\overline{\boldsymbol{X}}_{i-}\gamma_i\overline{\boldsymbol{x}}_i)$$

$$g_{3i}(\sigma_e^2, \sigma_v^2) = n_i^{-2}(\sigma_v^2 + \sigma_e^2 n_i^{-1})^{-3} h(\sigma_v^2, \sigma_e^2)$$

$h(\sigma_v^2, \sigma_e^2)$는 다음과 같이 주어진다.

$$h(\sigma_v^2, \sigma_e^2) = \sigma_e^4 var(\tilde{\sigma}_v^2) - 2\sigma_v^2\sigma_e^2 cov(\tilde{\sigma}_v^2, \tilde{\sigma}_e^2) + \sigma_v^4 var(\tilde{\sigma}_e^2) \quad 4\text{-}(13)$$

여기서, $var(\tilde{\sigma}_v^2) = 2n_*^{-2}[\nu_1^{-1}(n-p-\nu_1)(n-p)\sigma_e^4 + n_{**}\sigma_v^4 + 2n_*\sigma_v^2\sigma_e^2]$

$$var(\tilde{\sigma}_e^2) = 2\nu_1^{-1}\sigma_e^4$$

$$cov(\tilde{\sigma}_v^2, \tilde{\sigma}_e^2) = -2n_*^{-1}\nu_1^{-1}(n-p-\nu_1)\sigma_e^4$$

$$n_{**} = \sum_{i=1}^{m} n_i^2(1 - n_i\overline{\boldsymbol{x}}'_i(\sum_{i=1}^{m} \boldsymbol{x}'_i \boldsymbol{V}_{i^{-1}}\boldsymbol{x}_i)^{-1}\overline{\boldsymbol{x}}_i) + tr[(\sum_{i=1}^{m} \boldsymbol{x}'_i \boldsymbol{V}_{i^{-1}}\boldsymbol{x}_i)^{-1} n_i^2 \overline{\boldsymbol{x}}_i \overline{\boldsymbol{x}'}_i]$$

$MSE(\theta_i)$의 추정량은 다음과 같다.

$$mse(\hat{\theta}_i) = g_{1i}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) + g_{2i}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) + 2g_{3i}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) \quad 4\text{-}(14)$$

이 추정량은 최적비편향모형이다. EBLUP 추정량 $\hat{\theta}_i$은 완벽하게 모형 기반이고, 조사 가중치 $\tilde{w}_{ij}$를 사용하지 않는다. 따라서, 일반적으로 n_i증가에 의한 모형 일치성은 없다. 조사자는 실제조사에서 일반적으로 표본크기 n_i가 적더라도 모형일

치성을 갖기를 원한다. 그러나, 적절한 모형일치성을 얻기 위해서는 지역에서 n_i가 적당히 크기를 가져야만 한다.

(2) Prasad-Rao EBLUP 추정량

소지역 평균 θ_i의 Prasad-Rao EBLUP 추정량은 모형일치성을 갖는 추정량 $\bar{y}_{iw}$를 사용한다. 이 추정량은 BLUP 이론에 지역별모형 4-(6)에 기반한 소지역 평균 $\theta_i = \overline{\boldsymbol{X}}'_i\boldsymbol{\beta} + v_i$의 BLUP 추정량을 기본으로 하고, 다음과 같이 주어진다.

$$\tilde{\theta}_{i,aw} = \gamma_{iw}\bar{y}_{iw} + (\overline{\boldsymbol{X}}_i - \gamma_{iw}\bar{\boldsymbol{x}}_{iw})'\tilde{\boldsymbol{\beta}}_{aw} \qquad 4\text{-}(15)$$

여기서, $\gamma_{iw} = \sigma_v^2/(\sigma_v^2 + \sigma_e^2\delta_{iw})$

$\tilde{\boldsymbol{\beta}}_{aw}$ 는 다음과 같다.

$$\tilde{\boldsymbol{\beta}}_{aw} = (\sum_{i=1}^{m}\gamma_{iw}\bar{\boldsymbol{x}}_{iw}\bar{\boldsymbol{x}}'_{iw})^{-1}(\sum_{i=1}^{m}\gamma_{iw}\bar{\boldsymbol{x}}_{iw}\bar{y}_{iw}) \equiv \tilde{\boldsymbol{\beta}}_{aw}(\sigma_v^2, \sigma_e^2) \qquad 4\text{-}(16)$$

σ_v^2 와 σ_e^2 주어지고, $\tilde{\boldsymbol{\beta}}_{aw}$의 기댓값은 $E(\tilde{\boldsymbol{\beta}}_{iw}) = \boldsymbol{\beta}$이고, $\tilde{\boldsymbol{\beta}}_{aw}$의 분산은 다음식으로 주어진다.

$$var(\tilde{\boldsymbol{\beta}}_{aw}) = \sigma_v^2(\sum_{i=1}^{m} r_{iw}\bar{\boldsymbol{x}}_{iw}\bar{\boldsymbol{x}}'_{iw})^{-1} \equiv \boldsymbol{\Phi}_{aw} \qquad 4\text{-}(17)$$

지역별모형에 기반한 $\boldsymbol{\beta}$의 추정량 $\tilde{\boldsymbol{\beta}}_{aw}$는 단위별 모델에 비하여 주요효율 손실이 발생하기도 한다. 소지역평균의 추정에서 효율손실이 발생이 따를 수 있다.

주의할 것은 $\tilde{\theta}_{i,aw}$와 $\tilde{\boldsymbol{\beta}}_{aw}$는 알려지지 않은 σ_v^2 와 σ_e^2에 의존한다는 것이다. EBLLUP 추정에서 주어진 적합추정량 $\hat{\sigma}_v^2$ 와 $\hat{\sigma}_e^2$ 를 σ_v^2 와σ_e^2에 대신하여 4-(15) 4-(16)에 적용하면, 다음의 θ_i의 Prasad-Rao EBLUP 추정량을 얻을 수 있다.

$$\hat{\theta}_{i.aw} = \hat{\gamma}_{iw}\bar{y}_{iw} + (\bar{\boldsymbol{X}}_i - \hat{\gamma}_{iw}\bar{\boldsymbol{x}}_{iw})'\hat{\boldsymbol{\beta}}_{aw} \quad 4\text{-}(18)$$

여기서, $\hat{\gamma}_i = \hat{\sigma}_v^2/(\hat{\sigma}_v^2 + \hat{\sigma}_e^2\delta_{iw})$, $\hat{\boldsymbol{\beta}}_{aw} = \tilde{\boldsymbol{\beta}}_{aw}(\hat{\sigma}_v^2, \hat{\sigma}_e^2)$

이에 따라서 Prasad-Rao EBLUP 추정량 $\theta_{i.aw}$의 MSE 근사값은 다음식이 주어진다.

$$MSE(\theta_{i.aw}) \approx g_{1iw}(\sigma_v^2, \sigma_e^2) + g_{2i.aw}(\sigma_v^2, \sigma_e^2) + g_{3iw}(\sigma_v^2, \sigma_e^2) \quad 4\text{-}(19)$$

여기서, $g_{1iw}(\sigma_e^2, \sigma_v^2) = (1-\gamma_{iw})\sigma_e^2$

$$g_{2i.aw}(\sigma_e^2, \sigma_v^2) = (\bar{\boldsymbol{X}}_i - \gamma_{iw}\bar{\boldsymbol{x}}_{iw})'\boldsymbol{\Phi}_{aw}(\bar{\boldsymbol{X}}_i - \gamma_{iw}\bar{\boldsymbol{x}}_{iw})$$

$$g_{3iw}(\sigma_e^2, \sigma_v^2) = \gamma_{iw}(1-\gamma_{iw})^2\sigma_v^{-2}\sigma_e^{-2}h(\sigma_v^2, \sigma_e^2)$$

여기서 공분산행렬 $\boldsymbol{\Phi}_{aw}$는 식 4-(17)에서 ,$h(\sigma_v^2, \sigma_e^2)$은 식 4-(13)에서 주어진다. 그리고, $MSE(\theta_{i.aw})$ 최적비편향모형 추정치는 다음과 같이 주어진다.

$$mse(\hat{\theta}_{i.aw}) = g_{1iw}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) + g_{2i.aw}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) + 2g_{3iw}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) \quad 4\text{-}(20)$$

3) 제안하는 Pseudo-EBLUP 추정량

먼저 지역별모형 4-(6)에서 $\boldsymbol{\beta}$, σ_v^2 와 σ_e^2가 알려진다는 것을 가정하면, 소지역평균 θ_i의 BLUP 추정량은 다음과 같이 주어진다.

$$\tilde{\theta}_{iw} = \gamma_{iw}\bar{y}_{iw} + (\bar{\boldsymbol{X}}_i - \gamma_{iw}\bar{\boldsymbol{x}}_{iw})'\boldsymbol{\beta} \quad 4\text{-}(21)$$

BLUP 추정량 $\tilde{\theta}_{iw}$ $\boldsymbol{\beta}$, σ_v^2 와 σ_e^2에 의존한다. EBLUP에서 $\hat{\sigma}_v^2$ 와 $\hat{\sigma}_e^2$에 의한 단위별모형으로부터 σ_v^2 와 σ_e^2를 추정한다. 회귀계수 $\boldsymbol{\beta}$를 추정하기 위해, 먼저 v_i의

BLUP 추정량을 주어진 $\boldsymbol{\beta}$, σ_v^2 와 σ_e^2와 지역별모형 4-(6)에 의해 얻을 수 있다.

$$\tilde{v}_{iw}(\boldsymbol{\beta}, \sigma_v^2, \sigma_e^2) = r_{iw}(\bar{y}_{iw} - \bar{\boldsymbol{x}}'_{iw}\boldsymbol{\beta}) \qquad 4\text{-}(22)$$

그리고 나서 $\boldsymbol{\beta}$를 구하기 위한 다음의 조사 가중치 추정식을 풀면,

$$\sum_{i=1}^{m}\sum_{j=i}^{n_i}\tilde{w}_{ij}x_{ij}\{y_{ij} - x'_{ij}\boldsymbol{\beta} - \tilde{v}_{iw}(\boldsymbol{\beta}, \sigma_v^2, \sigma_e^2)\} = 0 \qquad 4\text{-}(23)$$

식 4-(23)로 부터 $\boldsymbol{\beta}$의 추정치를 얻을 수 있다.

$$\begin{aligned}\tilde{\boldsymbol{\beta}}_w &= \left\{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\tilde{w}_{ij}x_{ij}(x_{ij} - \gamma_{iw}\bar{x}_{iw})'\right\}^{-1}\left\{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\tilde{w}_{ij}(x_{ij} - \gamma_{iw}\bar{x}_{iw})y_{ij}\right\} \\ &\equiv \tilde{\boldsymbol{\beta}}_w(\sigma_v^2, \sigma_e^2)\end{aligned} \qquad 4\text{-}(24)$$

주어진, σ_v^2 와 σ_e^2와 $\tilde{\boldsymbol{\beta}}_w$는 $\boldsymbol{\beta}$의 모형비편향을 갖는다. 주의할 것은 $\tilde{\boldsymbol{\beta}}_w$는 단위별 모형 4-(4), 지역별모형4-(6)과 조사가중치 $\tilde{w}_{ij}$에 의해 얻었다는 것이다. 식 4-(24)에 기반하고 단위별모형 4-(4)에 의해 $\tilde{\boldsymbol{\beta}}_w$는 다음과 같다.

$$\tilde{\boldsymbol{\beta}}_w|\boldsymbol{\beta}, \sigma_v^2, \sigma_e^2 \sim N(\boldsymbol{\beta}, \boldsymbol{\Phi}_w) \qquad 4\text{-}(25)$$

공분산행렬 $\boldsymbol{\Phi}_w$는 다음과 같이 주어지고,

$$\begin{aligned}\boldsymbol{\Phi}_w = {} & \sigma_e^2\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}x_{ij}z'_{ij}\right)^{-1}\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}z_{ij}z'_{ij}\right)\left\{\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}x_{ij}z'_{ij}\right)^{-1}\right\}' \\ & + \sigma_v^2\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}x_{ij}z'_{ij}\right)^{-1}\left\{\sum_{i=1}^{m}\left(\sum_{j=1}^{n_i}z_{ij}\right)\left(\sum_{j=1}^{n_i}z_{ij}\right)'\right\}\left\{\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}x_{ij}z'_{ij}\right)^{-1}\right\}'\end{aligned} \qquad 4\text{-}(26)$$

$z_{ij} = \tilde{w}_{ij}(x_{ij} - \gamma_{iw}\bar{x}_{iw})$이다. 공분산 행렬 Φ_w은 σ_v^2 와 σ_e^2에 의존한다. 이제 식 4-(24)에 σ_v^2 와 σ_e^2 대신에 추정치 $\hat{\sigma}_v^2$ 와 $\hat{\sigma}_e^2$로 바꾸어 놓으면, $\hat{\boldsymbol{\beta}}_w$라는 $\boldsymbol{\beta}$의 조사가중치 추정량을 얻을 수 있다. 그리고 나서 $\boldsymbol{\beta}$, σ_v^2 와 σ_e^2 대신에 $\hat{\boldsymbol{\beta}}_w$, $\hat{\sigma}_v^2$ 와 $\hat{\sigma}_e^2$를 식 4-(21)에 바꿔어 놓으면, 소지역평균 θ_i의 새로운 Pseudo-EBLUP 추정량 $\hat{\theta}_{iw}$를 얻을 수 있다.

$$\hat{\theta}_{iw} = \hat{\gamma}_{iw}\bar{y}_{iw} + (\bar{X}_i - \hat{\gamma}_{iw}\bar{x}_{iw})'\hat{\boldsymbol{\beta}}_w \qquad 4\text{-}(27)$$

$\hat{\theta}_{iw}$를 다음과 같이 쓸 수도 있다.

$$\hat{\theta}_{iw} = \bar{X}'_i\hat{\boldsymbol{\beta}}_w + \hat{v}_{iw} \qquad 4\text{-}(28)$$

위 식에서, $\hat{v}_{iw} = \tilde{v}_{iw}(\hat{\boldsymbol{\beta}}_w, \sigma_v^2, \sigma_e^2)$ 이고 $\tilde{v}_{iw}$ 는 식 4-(22)에서 주어진다.

여기에서 n_i가 크게 됨에 따라서 제안하는 Pseudo-EBLUP 추정량 $\hat{\theta}_{iw}$는 설계 일치성을 갖는다. 그 이유는, 첫째로, $\gamma_{iw} \to 1 \quad \max j(\tilde{w}_{ij}) = O(n_i^{-1})$가 가정됨에 따라 $\gamma_{iw} \to 1$이 되기 때문이고, 둘째로, $\bar{y}_{iw}$는 $\bar{Y}_i$로 수렴하기 때문이며, 셋째로, $\bar{X}_i - \bar{x}_{iw}$는 0으로 수렴하기 때문이다. 한편, Prasad-Rao EBLUP $\hat{\theta}_{i.aw}$도 설계일치성을 가진다. 그러나, EBLUP 추정량 $\hat{\theta}_i$는 설계일치성을 갖지 않는다. n_i가 고정되어 있다고 하면, $m \to \infty$ 즉 소지역이 무한이 많아지게 되면 EBLUP 와 Pseudo-EBLUP 추정량은 BLUP 와 Pseudo-EBLUP 추정량과 유사하게 되는 경향이 있게 된다. 이는 $m \to \infty$함에 따라 $\hat{\sigma}_v^2$, $\hat{\sigma}_e^2$ 와 $\hat{\boldsymbol{\beta}}_{aw}(\hat{\boldsymbol{\beta}}_w)$가 모형일치성을 갖기 때문이다.

제안하는 Pseudo-EBLUP 추정량의 이점은 추정량 $\hat{\theta}_{iw}$가 i 상에서 합쳐질 때, 자동적으로 벤치마킹성질을 만족한다는 것이다. 가중치 $\tilde{w}_{ij}$가 알려진 모집단 합계

N_i에 일치되어 지는 것을 가정한다.

$$\sum_{j=1}^{n_i} \tilde{w}_{ij} = \tilde{w}_{i.} = N_i \qquad 4\text{-}(29)$$

$\sum_{i=1}^{m} N_i \hat{\theta}_{iw}$는 Y의 전체총합의 직접조사회귀 추정량이다.

$$\sum_{i=1}^{m} N_i \hat{\theta}_{iw} = \hat{Y}_w + (X - \hat{X}_w)' \beta_w \qquad 4\text{-}(30)$$

여기서, $\hat{Y}_w = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \tilde{w}_{ij} y_{ij} = \sum_{i=1}^{m} \tilde{w}_{i.} \bar{y}_{iw}$

$$\hat{X}_w = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \tilde{w}_{ij} x_{ij}$$

$\hat{Y}_w$와 $\hat{X}_w$는 각각 Y와 X의 합의 직접 추정량이 된다. 이 결과에서 볼 수 있는 것은, 먼저 $\boldsymbol{\beta}$가 $\hat{\boldsymbol{\beta}}_w$에 의해, $\tilde{\mathbf{v}}_{iw}(\boldsymbol{\beta}, \sigma_v^2, \sigma_e^2)$가 $\hat{\mathbf{v}}_{iw}$에 의해 추정되어 진다는 것이다. 식 4-(23)에 $\boldsymbol{\beta}$와 $\tilde{\mathbf{v}}_{iw}(\boldsymbol{\beta}, \sigma_v^2, \sigma_e^2)$ 를 대신하여 $\hat{\boldsymbol{\beta}}_w$와 $\hat{\mathbf{v}}_{iw}$ 적용하고 절편 $(x_{ij1} = 1)$을 주면

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} \tilde{w}_{ij} (y_{ij} - x'_{ij} \hat{\boldsymbol{\beta}}_w - \hat{\mathbf{v}}_{iw}) = 0 \qquad 4\text{-}(31)$$

$$\sum_{i=1}^{m} N_i \hat{\mathbf{v}}_{iw} = \hat{Y}_w - \hat{X}_w' \hat{\boldsymbol{\beta}}_w \qquad 4\text{-}(32)$$

로 표현이 가능하다. 식 4-(28)과 식 4-(30)로부터 다음식을 얻을 수 있다.

$$\sum_{i=1}^{m} N_i \hat{\theta}_{iw} = X \hat{\boldsymbol{\beta}}_w + \sum_{i=1}^{m} N_i \hat{\mathbf{v}}_{iw} = \hat{Y}_w + (X - \hat{X}_w)' \hat{\boldsymbol{\beta}}_w \qquad 4\text{-}(33)$$

제안하는 Pseudo-EBLUP 추정량 $\hat{\theta}_{iw}$는 어떠한 수정이 없이 벤치마킹성질을 만족한다. EBLUP 추정량 $\hat{\theta}_i$와 Prasad-Rao EBLUP 추정량 $\hat{\theta}_{i.aw}$는 제안하는 Pseudo-EBLUP 추정량 $\hat{\theta}_{iw}$과는 다르게 Y의 직접조사회귀추정량을 위한 벤치마킹을 하지 않는다. $\hat{\theta}_i$ 와 $\hat{\theta}_{i.aw}$ 은 벤치마킹성질을 만족하기 위해서, 예를들면, 단순비율 벤치마킹의 사용과 같은 조정이 필요하다. 그러나 벤치마킹에 의한 MSE 추정량은 수정된 추정량에 따른 변이성질을 가지지 않는다. 따라서, θ_{iw} MSE 추정량으로 다음식을 사용한다.

$$MSE(\tilde{\theta}_{iw}) \approx g_{1iw}(\sigma_v^2, \sigma_e^2) + g_{2iw}(\sigma_v^2, \sigma_e^2) + g_{3iw}(\sigma_v^2, \sigma_e^2) \qquad 4\text{-}(34)$$

여기서, $g_{2iw}(\sigma_v^2, \sigma_e^2) = (\overline{X}_i - \gamma_{iw}\overline{x}_{iw})'\Phi_w(\overline{X}_i - \gamma_{iw}\overline{x}_{iw})$

$g_{1iw}(\sigma_v^2, \sigma_e^2)$ 와 $g_{3iw}(\sigma_v^2, \sigma_e^2)$ 은 식 4-(19)에서와 같다. 또한 Φ_w은 식 4-(26)에 주어져 있으며, MSE 근사치의 추정량은 다음과 같이 얻어질 수 있다.

$$mse(\hat{\theta}_{iw}) \approx g_{1iw}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) + g_{2iw}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) + 2g_{3iw}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) \qquad 4\text{-}(35)$$

Ⅲ 활용과 결론

<표 1> 2008년도 서귀포시 조생밀감 표본 조사결과

구분	동별 재배면적 (ha) N_i	표본수 n_i	표본			동별 평균	
			1주당 실중량(kg) y_{ij}	10a당주수 (개) x_{ij1}	열매수 (개) x_{ij2}	10a당주수 (개) $x_{i1(p)}$	열매수 (개) $x_{i2(p)}$
송산	179	1	33.0	132	275	107.85	264.4
효돈	356	3	2.7	81	25	72.38	489.5
			104.8	69	981		
			74.8	105	702		
영천	668	5	60.4	107	618	108.23	688.2
			72.0	52	692		
			51.7	104	458		
			69.9	119	817		
			77.1	116	915		
동홍	218	2	64.2	58	437	109.50	365.25
			62.1	107	391		
서홍	278	2	46.1	107	422	109.82	399.43
			40.6	83	379		
대륜	518	4	90.3	111	806	86.68	949.57
			42.8	155	325		
			131.9	52	1301		
			91.0	79	756		
대천	617	5	51.8	119	754	111.89	585.88
			84.8	83	728		
			87.9	101	726		
			28.6	129	354		
			56.9	102	471		
중문	792	6	51.5	129	546	124.40	467.43
			61.3	119	713		
			52.9	111	415		
			38.8	112	345		
			34.7	119	302		
			27.1	108	671		
예례	657	5	67.1	101	480	97.15	424.27
			17.0	113	618		
			11.8	92	329		
			14.6	85	337		
			22.0	95	809		

<표 1>은 EBLUP 추정량과 제안하는 Pseudo-EBLUP 추정량에 대한 실례를 적용하기 위해 2008년도 실제 서귀포시 조생밀감 조사 자료를 통헤 서귀포시 지역 생산량을 추정하였다. 어려움 속에서도 추정의 방법과 정확성을 기하려는 연구목적에 동의하면서 2008년도 밀감생산조사 자료를 이용할 수 있도록 편의를 주신 서귀포시 농업기술원에 감사를 드립니다. 서귀포시 10개동 중 천지동은 실제조사 농가가 없어 표본크기가 영이 되는 경우가 발생하였는데 대여력(borrowing stength)에 의해 추정 가능한 부분이므로 천지동을 송산동에 포함하여 9개 동을 소지역으로 선정하고 여기에서 농업기술원에서 조사한 총 100농가 중 단순임의추출방식으로 다시 33개를 추출하여 조사하는 방식을 취하였다. 각 지역의 표본크기는 서귀포시 총 재배면적에 따라 각 소지역에 표본크기 $n_i\,(i=1,\cdots,9)$는 지역별로 1개에서 6개의 범위에 있고, 지역별재배면적 N_i(모집단 크기)는 179ha에서 792 ha 범위에 있다.

지역내에서의 표본은 단순임의추출을 통해서 추출한다. 지역 i에서 조사가중치는 $\widetilde{w}_{ij}=N_i/n_i$이고 $w_{ij}=1/n_i$. 표본의 모형은 다음과 같다.

$$y_{ij}=\beta_0+x_{ij1}\beta_1+x_{ij2}\beta_2+v_i+e_{ij},\quad j=1,\cdots,n_i,\quad i=1,\cdots,9$$

y_{ij}는 i동의 j번째 농가의 감귤나무 1주당 생산량이고, x_{ij1}은 i동의 j번째 농가의 10a 당 감귤나무 주수, x_{ij1}은 i동의 j번째 농가의 감귤나무 1주당 열매수이다. 9개동에 대해 제안하는 Pseudo-EBLUP 추정량을 통해 총계 추정치를 구할 수 있다. 이제 추정량을 구하기 위해 다음과 같은 과정을 거치게 된다.

EBLUP 추정량 : (1) 오차 분산 $\hat{\sigma}_v^2$와 $\hat{\sigma}_e^2$를 구한다. (2) 블록대각선행렬에 의한 $\mathbf{V}$를 구한다. (3) 회귀계수 $\boldsymbol{\beta}$의 추정량 $\hat{\boldsymbol{\beta}}$과 그 표준편차를 구한다. (4) θ의 추정량 $\hat{\theta}$를 구한다. (5) 추정량 $\hat{\theta}$의 MSE를 구한다.

Parasad-Rao EBLUP 추정량 : (1) 오차 분산 $\hat{\sigma}_v^2$와 $\hat{\sigma}_e^2$를 구한다. (2) 회귀계수 $\boldsymbol{\beta}$의 추정량 $\hat{\boldsymbol{\beta}}_{aw}$과 그 표준편차를 구한다. (3) θ의 추정량 $\hat{\theta}_{i.aw}$를 구한다. (4) 추정량 $\hat{\theta}_{i.aw}$의 MSE를 구한다.

제안하는 Pseudo-EBLUP 추정량 : (1) 오차 분산 $\hat{\sigma}_v^2$와 $\hat{\sigma}_e^2$를 구한다. (2) 가중치에 의한 회귀계수 $\boldsymbol{\beta}$의 추정량 $\hat{\boldsymbol{\beta}}_{iw}$과 그 표준편차를 구한다. (3) θ의 추정량 $\hat{\theta}_{iw}$를 구한다. (4) 추정량 $\hat{\theta}_{iw}$의 MSE를 구한다. (5) 총계추정량을 구한다.

오차 분산 $\hat{\sigma}_v^2$와 $\hat{\sigma}_e^2$는 모든 추정량에서 공통으로 사용되어 지고, 그 값은 $\hat{\sigma}_v^2 = 113.1007$ 과 $\hat{\sigma}_e^2 = 273.7772$로 주어진다.

<표 2> 회귀계수에 대한 추정식과 그 특징

구분		내용
EBLUP 추정량 $\hat{\boldsymbol{\beta}}$	추정식	$\tilde{\boldsymbol{\beta}} = (\sum_{i=1}^{m} \boldsymbol{x}'_i \boldsymbol{V}_i^{-1} \boldsymbol{x}_i)^{-1} (\sum_{i=1}^{m} \boldsymbol{x}'_i \boldsymbol{V}_i^{-1} \boldsymbol{y}_i) \equiv \tilde{\boldsymbol{\beta}}(\sigma_e^2, \sigma_v^2)$
	분산	$var(\hat{\boldsymbol{\beta}}) = (\sum_{i=1}^{m} \boldsymbol{x}'_i \boldsymbol{V}_i^{-1} \boldsymbol{x}_i)^{-1}$
	특징	표본자료, 블록 대각 공분산 행렬 $\boldsymbol{V}$ 사용
Parasad-Rao EBLUP 추정량 $\hat{\boldsymbol{\beta}}_{aw}$	추정식	$\tilde{\boldsymbol{\beta}}_{aw} = (\sum_{i=1}^{m} \gamma_{iw} \bar{\boldsymbol{x}}_{iw} \bar{\boldsymbol{x}}'_{iw})^{-1} (\sum_{i=1}^{m} \gamma_{iw} \bar{\boldsymbol{x}}_{iw} \bar{y}_{iw}) \equiv \tilde{\boldsymbol{\beta}}_{aw}(\sigma_v^2, \sigma_e^2)$
	분산	$var(\tilde{\boldsymbol{\beta}}_{aw}) = \sigma_v^2 (\sum_{i=1}^{m} r_{iw} \bar{\boldsymbol{x}}_{iw} \bar{\boldsymbol{x}}'_{iw})^{-1} \equiv \boldsymbol{\Phi}_{aw}$
	특징	표본자료의 평균값, r_{iw} 사용
제안하는 Pseudo-EBLUP 추정량 $\hat{\boldsymbol{\beta}}_{iw}$	추정식	$\tilde{\boldsymbol{\beta}}_w = \left\{\sum_{i=1}^{m} \sum_{j=1}^{n_i} \tilde{w}_{ij} x_{ij} (x_{ij} - \gamma_{iw} \bar{x}_{iw})'\right\}^{-1} \left\{\sum_{i=1}^{m} \sum_{j=1}^{n_i} \tilde{w}_{ij} (x_{ij} - \gamma_{iw} \bar{x}_{iw}) y_{ij}\right\} \equiv \tilde{\boldsymbol{\beta}}_w(\sigma_v^2, \sigma_e^2)$
	분산	$var(\tilde{\boldsymbol{\beta}}_w) = \sigma_e^2 (\sum_{i=1}^{m} \sum_{j=1}^{n_i} x_{ij} z'_{ij})^{-1} (\sum_{i=1}^{m} \sum_{j=1}^{n_i} z_{ij} z'_{ij}) \left\{(\sum_{i=1}^{m} \sum_{j=1}^{n_i} x_{ij} z'_{ij})^{-1}\right\}' + \sigma_v^2 (\sum_{i=1}^{m} \sum_{j=1}^{n_i} x_{ij} z'_{ij})^{-1} \left\{\sum_{i=1}^{m} (\sum_{j=1}^{n_i} z_{ij})(\sum_{j=1}^{n_i} z_{ij})'\right\} \left\{(\sum_{i=1}^{m} \sum_{j=1}^{n_i} x_{ij} z'_{ij})^{-1}\right\}' = \boldsymbol{\Phi}_w$
	특징	총계추정치를 위한 가중치를 이용한 $z_{ij} = \tilde{w}_{ij}(x_{ij} - \gamma_{iw} \bar{x}_{iw})$와 표본자료, r_{iw}사용

다음으로 회귀계수 β를 추정값을 구하기 위해 각 추정량에 대한 추정식을 정리하였다. <표 2>는 회귀계수 β를 추정값 추정식과 그에 따른 분산 그리고 추정량별 특징을 설명한 표이다.

이제 각 추정량을 구하고 추정량에 대한 표준편차를 구하였다. <표 3>은 회귀계수 β에 대한 추정값과 표준편차를 구한 결과이다.

<표 3> 회귀계수에 대한 추정치와 표준편차

구분		EBLUP 추정량 $\hat{\beta}$	Parasad-Rao EBLUP 추정량 $\hat{\beta}_{aw}$	제안하는 Pseudo-EBLUP 추정량 $\hat{\beta}_{iw}$
추정치	β_0	36.998	23.607	36.787
	β_1	-0.216	-0.206	-0.212
	β_2	0.073	0.095	0.072
표준편차	β_0	18.496	47.254	18.545
	β_1	0.142	0.385	0.142
	β_2	0.013	0.035	0.013

<표 4> $\hat{\theta}_{iw}$를 구하기 위한 주요변수 계산 결과

구분	동별 재배면적(ha) N_i	$\hat{\gamma}_{iw}$	표본평균			동별 평균	
			1주당 실중량(kg) $\bar{y}_{iw}$	10a당주수(개) $\bar{x}_{1iw}$	1주당열매수(개) $\bar{x}_{2iw}$	10a당주수(개) $\bar{X}_{1i}$	1주당열매수(개) $\bar{X}_{2i}$
송산	179	0.29	33.00	132.00	275.00	107.85	264.4
효돈	356	0.55	60.70	85.00	569.33	72.38	489.50
영천	668	0.67	66.22	99.45	700.00	108.23	688.20
동홍	218	0.45	63.15	82.12	414.00	109.50	365.25
서홍	278	0.45	43.35	95.25	400.50	109.82	399.43
대륜	518	0.62	89.00	99.00	797.00	86.68	949.57
대천	617	0.67	62.00	106.80	566.20	111.89	585.88
중문	792	0.71	44.38	116.38	498.67	121.40	467.43
예례	657	0.67	26.50	96.90	514.60	97.15	424.27

소지역 평균 $\theta_i = \beta_0 + \overline{X}_{i1}\beta_1 + \overline{X}_{i2}\beta_2 + v_i, \quad i = 1, \cdots, 9$이고, $\overline{X}_{i1}$와 $\overline{X}_{i2}$는 각각 i번재 소지역의 10a 당 감귤나무 주수 모집단 평균, i번재 소지역의 감귤나무 1주당 열매수 모집단 평균이 된다. θ_i의 추정량을 구할 수 있다. <표 4>는 추정식을 구하기 위해 식에 적용되는 주요 변수들에 대해 계산한 결과이다.

이제 <표 3>의 회귀계수 $\boldsymbol{\beta}$의 추정치와 <표 4>에서 구한 주요 변수들을 통해 θ_i 추정량을 구할 수 있다.

일례로 송산지역의 1주당 실중량을 제안하는 Pseudo-EBLUP 추정식에 적용하여 실제로 구해보면 추정량 $\hat{\theta}_{1w}$은

$$\hat{\theta}_{1w} = \hat{\gamma}_{1w}\bar{y}_{1w} + (\overline{X}_{11} - \hat{\gamma}_{1w}\bar{x}_{11w})'\hat{\boldsymbol{\beta}}_w = 0.29 \times 33.00 +$$

$$(1 - 0.29, 107.85 - 0.29 \times 132.00, 264.40 - 0.29 \times 275.00)\begin{pmatrix} 36.79 \\ -0.21 \\ 0.07 \end{pmatrix}$$

$$= 34.26$$

이 된다. 계산 방식은 다른 추정식에서도 같게 적용된다.

이제 θ_i의 추정량에 대한 MSE 값을 구하여 보자. MSE값의 각 추정값 mse는 공통적으로 같은 형태를 가지게 된다.

$mse(\hat{\theta}_i) = g_{1i}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) + g_{2i}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) + 2g_{3i}(\hat{\sigma}_v^2, \hat{\sigma}_e^2)$이고 여기서 각 추정값들 중 차이가 있는 부분은 회귀계수 $\boldsymbol{\beta}$ 추정치의 분산의 영향을 받게 되는 $g_{2i}(\hat{\sigma}_v^2, \hat{\sigma}_e^2)$만 차이가 있게 된다. <표 5>는 mse 추정식 내용이다.

<표 5> 회귀계수에 대한 mse 추정식 내용

구분	EBLUP 추정량 $\hat{\boldsymbol{\beta}}$	Parasad-Rao EBLUP 추정량 $\hat{\boldsymbol{\beta}}_{aw}$	제안하는 Pseudo-EBLUP 추정량 $\hat{\boldsymbol{\beta}}_{iw}$
$g_1(\hat{\sigma}_v^2, \hat{\sigma}_e^2)$	$(1-\gamma_{iw})\sigma_e^2$		
$g_2(\hat{\sigma}_v^2, \hat{\sigma}_e^2)$	$g_2(\sigma_e^2, \sigma_v^2) = (\overline{\boldsymbol{X}}_{i-}\gamma_{iw}\overline{\boldsymbol{x}}_i)'$분산$(\overline{\boldsymbol{X}}_{i-}\gamma_{iw}\overline{\boldsymbol{x}}_i)$		
	$var(\hat{\boldsymbol{\beta}})$	Φ_{aw}	Φ_w
$g_3(\hat{\sigma}_v^2, \hat{\sigma}_e^2)$	$g_3(\sigma_e^2, \sigma_v^2) = \gamma_{iw}(1-\gamma_{iw})^2\sigma_v^{-2}\sigma_e^{-2}h(\sigma_v^2, \sigma_e^2)$		

따라서, <표 3>,<표 4>,<표 5>의 결과에 따라 θ_i의 추정치와 표준편차(mse)를 계산 하면 다음과 같이 계산할 수 있다.

<표 6> θ_i의 추정치 및 표준편차 계산결과

구분	표본수	EBLUP $\hat{\theta}_i$		Parasad-Rao EBLUP $\hat{\theta}_{aw}$		제안하는 Pseudo-EBLUP $\hat{\theta}_{iw}$	
		추정량	표준편차	추정량	표준편차	추정량	표준편차
송산	1	34.27	11.17	29.63	13.13	34.26	11.18
효돈	3	57.38	9.20	55.57	12.10	57.25	9.21
영천	5	63.55	7.34	64.46	8.11	63.48	7.34
동홍	2	46.15	9.93	43.56	11.92	46.20	9.94
서홍	2	41.35	9.76	39.59	10.72	41.34	9.77
대륜	4	96.97	8.46	102.33	11.88	96.63	8.46
대천	5	60.10	7.29	60.72	7.81	60.03	7.30
중문	6	42.11	6.81	41.17	7.62	42.12	6.82
예례	5	28.70	7.34	26.40	8.10	28.70	7.34

이제 제안하는 Pseudo-EBLUP $\hat{\theta}_{iw}$를 이용하여 각 지역별 총생산량에 대한 조사추정치를 구하였다. 우선 송산 지역 총 179ha 중 10a 당 1주 실중량 총합에 대해 구해보면, $N_1\hat{\theta}_{1w} = 61.32\,(ton)$이다. 따라서, 송산 지역 총 면적 179ha 전체에 대한 생산량 총합은 $N_1\hat{\theta}_{1w}$에 10a 당 주수(송산지역 평균)를 곱해주면 구할 수 있게 된다.

$$10a\text{당 주수(동평균)} \times N_1\hat{\theta}_{1w} = 107.85 \times 61.32 = 6613.90\,(ton)$$

그렇게 되면 다른 지역별 총생산량에 대해서 구할 수 있게 된다.

<표 7>은 지역별 총생산량에 대한 조사추정치를 나타낸다.

<표 7> 지역별 총생산량에 대한 조사추정치와 표준편차

지역	재배면적 (ha) N_i	표본수 n_i	Pseudo-EBLUP (kg) 10a당 1주중량 $\hat{\theta}_{iw}$	s.e	$N_i\hat{\theta}_{iw}$ (ton)	$N_i\hat{\theta}_{iw}$×10a당 주수 (동평균) 지역별총생산량 (ton)
송산	179	1	34.26	11.18	61.32	6613.90
효돈	356	3	57.25	9.21	203.82	14751.32
영천	668	5	63.48	7.34	424.05	45893.33
동흥	218	2	46.20	9.94	100.72	11028.46
서흥	278	2	41.34	9.77	114.92	12621.00
대륜	518	4	96.63	8.46	500.56	43386.96
대천	617	5	60.03	7.30	370.40	41442.70
중문	792	6	42.12	6.82	333.62	40499.74
예래	657	5	28.70	7.34	188.57	18318.98
계	4,283	33			2297.99	234556

. 최종적으로 서귀포시의 총 감귤생산량은 234,556(ton)으로 추정할 수 있다.

<표 8>은 현장조사를 통해 다음과 같은 사항에 대해 조사하고 그 결과를 바탕으로 생산예상량은 추정한 결과를 보여주고 있다.

<표 8> 현장조사결과에 따른 생산량 예측자료

구분	주당열매수(개)	수확시과실횡경(mm)	수확시예측과중(g)	10a당주수㈜	10a당수량(kg)	재배면적 (ha)	생산예상량(톤)
제주시	615	63.0	102.9	97	3,930	6,510	255,845
서귀포시	516	64.3	109.0	95	3,421	11,920	407,728
서귀포						4,283	146,502
도전체						18,430	663,574

서귀포시 총 감귤생산량은 146,502(ton)으로 계산결과에 의한 추정치와 많은 차이가 있으나 그 이유는 현장조사결과에 따른 생산량 예측시 다음식을 사용하여 추정하였기 때문이다.

10a당 생산량 = (주당열매수×수확시예측과중×10a당주수)×64%

따라서, 계산된 추정치에 같은 방법으로 64%를 적용하면
총 생산량 추정치는 234,556 × 0.64 = 150,116(ton)로 조사 결과 추정치와 비슷한 결과를 얻을 수 있다.

현장조사를 통한 예측은 총 100개의 표본에 의한 조사를 통해 결과가 얻어진 반면, 계산된 추정치는 33개의 표본만으로 비슷한 결과를 얻어 냈다는 점에서 제안하는 Pseudo-EBLUP 추정량의 강점이 있다고 할 수 있다. 특히 <표 6>에서 제시하는 것처럼 기존의 Parasad-Rao EBLUP보다도 제안하는 Pseudo-EBLUP는 소지역 추정에 있어서 각 지역당 표준편차가 현저히 감소한다는 점을 감안한다면 소지역 생산량의 추정에서 괄목할만한 정도의 효과를 나타낼 수 있다고 볼 수 있다.

특히, 연구결과가 실제 조사에 활용될 경우에 표본설계 과정에서 조사비용을 줄일 수 있는 방법이 될 것으로 예상된다. 하지만 여기에서 주의해야 할 점은 관심되는 주요 변수에 대응하는 보조변수에 대한 정확한 정보를 획득할 수 있어야 하는바 본 조사에서는 첫째, 조사농가당 재식거리에 대한 정확한 10a당 주수의 조사와 함께, 둘째는 1주당 실중량의 예측을 위하여 회귀모형의 선정에서 최적의 방법을 고려해야 한다는 것이다.

Ⅳ 참고문헌

[1] 제주특별자치도 농업기술원 현장조사결과 자료.(2008).

[2] 김익찬(2006). 소지역 추정량과 그 응용, Journal of the Korean Data Analysis Society. Vol 8. No 1. February 2006, 205-214.

[3] 안정용, 김대경, 최경호 (2005). 지역통계의 문제점과 개선방안, Jornal of the Korean Data Analysis Society, Vol. 7, No. 6, 2037-2047.

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

[5] 신민웅, 최기철 (2003). 집락추출시 최적 표본 추출, Jouranl of the Korean Data Analysis Society, Vol. 5, No. 3, 591-599.

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

[7] 이계오 (2002). 소지역추정법에 의한 시군구 실업 통계 개발 최종보고서, 통계청.

[8] 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.

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

[10] 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.

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