응용통계연구 제 15권 1호, 2002년, pp. 119-128

반복조사에서 소지역자료의 베이지안 분석

김달호[1] 김남희[2]

요 약

Rao와 Yu(1994)는 소지역 추정(small area estimation) 문제를 해결하기 위한 방법으로 추정 시점의 인접지역 정보 등 보조정보와 과거의 표본조사 결과를 모두 이용하는 모형과 그 모형으로 부터 경험적최량선형비편향추정량(Empirical Best Linear Unbiased Predictor)을 제안하였다. 본 논문에서는 Rao와 Yu의 모형에서 미지의 모수에 대한 사전확률분포를 가정한 계층적 베이즈 추정량을 제안하고, 이를 미국의 주별 4인가족 소득추정문제에 적용하여 그 효율을 미국의 Census Bureau에서 사용하고 있는 경험적 베이즈추정량 및 이전에 제안된 다른 추정량들과 비교하였다.

주요용어 : 소지역추정; 계층적 베이즈모형; 깁스샘플러; 반복조사.

1. 서론

소지역추정 문제란 넓은 지역을 대상으로 하는 조사에서, 표본설계 당시에는 고려되지 않은 하부단위 즉, 소지역의 특성을 파악코자 할 때, 표본오차와 변동계수 등이 허용치 이상으로 커짐으로 소지역에 대한 표본조사 결과를 신뢰할수 없는 상태를 뜻한다. 이러한 문제를 해결하기 위한 방법으로 표본조사 결과 뿐만 아니라 인접지역의 정보 혹은 과거의 정보를 모두 이용하는 추정모형을 생각할수 있다. Rao와 Yu(1994)는 이러한 모형하에서 최량선형비편향추정량을 제안하였다. 본 논문에서는 Rao와 Yu의 모형에서 미지의 모수에 대한 사전확률분포를 가정한 계층적 베이즈모형을 제안하였다. 소지역추정을 위하여 경험적 베이즈 추정방법과 계층적 베이즈 추정방법은 가장 많이 쓰이는 방법이다. 실제로 미국 센서스국에서는 주별 4인가족 소득을 추정하기 위하여 경험적 베이즈 추정량을 사용하고 있다. 이 두방법은 모수에 대한 불확실성을 가정한다는 점에서는 공통적이나, 계층적 베이즈 추정에서는 초모수(hyperparameter)에 대해서 분포를 적용하지만 경험적 베이즈추정에서는 최대우도추정법, 적률방법 등으로 이를 추정한다. 점추정에 있어서는 이들 방법의 결과가 비슷하지만 추정량의 분산 추정에 있어서는 계층적 베이즈 추정방법이 더 우세한데, 이는 경험적 베이즈 추정방법은 초모수를 추정함으로, 사후분산이 상당히 과소 추정되는 경향이 있기 때문이다. 계층적 베이즈 추정방법은 고차원의 수치적분을 풀어야 하는 계산문제를 안고 있지만, 근래에 들어서 깁스샘플러(Gibbs Sampler) 방법의 개발로 비교적 간단히 해결될 수 있다. (Gelfand와 Smith(1990) 참조.)

1) (702-701)대구광역시 북구 산격동 1370번지, 경북대학교 자연과학대학 통계학과, 부교수
E-mail: dalkim@knu.ac.kr
2) (702-702)대구광역시 북구 산격동 1445-3, 경북도청 기획관실
E-mail: jungdwn@hanmail.net

2절에서는 분석에 이용되는 모형을 소개하며, 3절에서는 Gelman, Meng 및 Stern(1996)에 의해 제안된 모형 적합성 판정의 측정기준을 소개한다. 4절에서는 소지역추정문제로 잘 알려진 미국의 주별 4인가족 소득추정을 위하여 이 논문에서 제안된 계층적 베이즈 추정량을 적용하였다. 현재 미국의 Census Bureau에서는 4인가족 소득추정을 위하여 경험적 베이즈 추정방법을 쓰고 있으며, 여러 논문에서 동일한 문제 해결을 위해 다양한 추정량들을 제시하였다. 상대적인 추정효율을 알아보기 위하여 본 논문의 추정량과 미국의 Census Bureau 및 다른 논문의 추정량을 함께 비교하여 보았다.

2. 현황자료와 시계열자료를 결합한 계층적 베이즈 모형

과거의 자료로부터 정보를 빌어오기 위하여 Rao와 Yu(1994)는 Fay와 Herriot(1979)의 모형을 시계열 모형으로 확장하였다. 이 모형은 다음과 같다.

$$\begin{aligned} y_{it} &= \boldsymbol{x}_{it}^T\boldsymbol{\beta} + v_i + u_{it} + e_{it}, \\ u_{it} &= \rho u_{i,t-1} + \eta_{it},\ |\rho| < 1, \\ & i = 1, ..., m,\ t = 1, ..., T, \end{aligned} \tag{1}$$

여기서 v_i, η_{it}, e_{it}는 상호 독립이며, $v_i \overset{ind}{\sim} N(0, \sigma_v^2)$, $\eta_{it} \overset{ind}{\sim} N(0, \sigma_\eta^2)$, $\boldsymbol{e} = (e_{it}) \overset{ind}{\sim} N(\boldsymbol{0}, \boldsymbol{\Sigma})$, 그리고 $\boldsymbol{\Sigma} = \text{Block Diag}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_m)$와 같다.

모형 (1)은 다음과 같은 행렬식으로 표현된다.

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{v} + \boldsymbol{u} + \boldsymbol{e} \\ \boldsymbol{X} &= (\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_m^T)^T,\ \boldsymbol{X}_i^T = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT}) \\ \boldsymbol{Z} &= \boldsymbol{I}_m \otimes \boldsymbol{1}_T,\quad \boldsymbol{v} = (v_1, \ldots, v_m)^T \\ \boldsymbol{u} &= (\boldsymbol{u}_1^T, \ldots, \boldsymbol{u}_m^T)^T,\ \boldsymbol{u}_i^T = (u_{i1}, \ldots, u_{iT}) \\ \boldsymbol{e} &= (\boldsymbol{e}_1^T, \ldots, \boldsymbol{e}_m^T)^T,\ \boldsymbol{e}_i^T = (e_{i1}, \ldots, e_{iT}), \end{aligned}$$

여기서 $\boldsymbol{1}_T$은 1로 구성된 T-벡터이며, $\boldsymbol{I}_m$는 m차 항등행렬이며 $\otimes$는 직접곱(direct product)이다. 각 변수들의 평균과 분산 행렬은 다음과 같다.

$$\begin{aligned} E(\boldsymbol{v}) &= \boldsymbol{0},\quad Cov(\boldsymbol{v}) = {\sigma_v}^2\boldsymbol{I}_m \\ E(\boldsymbol{u}) &= \boldsymbol{0},\quad Cov(\boldsymbol{u}) = \sigma^2\boldsymbol{I}_m \otimes \boldsymbol{\Gamma} = \sigma^2\boldsymbol{R} \\ E(\boldsymbol{e}) &= \boldsymbol{0},\quad Cov(\boldsymbol{e}) = \boldsymbol{\Sigma} = \text{Block Diag}(\boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\Sigma}_m) \\ Cov(\boldsymbol{y}) &= \boldsymbol{V} = \boldsymbol{\Sigma} + \sigma^2\boldsymbol{R} + {\sigma_v}^2\boldsymbol{Z}\boldsymbol{Z}^T \\ &= \text{Block Diag}(\boldsymbol{\Sigma}_i + \sigma^2\boldsymbol{\Gamma} + {\sigma_v}^2\boldsymbol{J}_T). \end{aligned}$$

여기서 $\boldsymbol{J}_T = \boldsymbol{1}_T\boldsymbol{1}_T^T$, 그리고, $\boldsymbol{v}$, $\boldsymbol{u}$, $\boldsymbol{e}$는 상호 독립이며, $\boldsymbol{\Gamma}$는 T × T 행렬로서, (i, j) 요소는

$\rho^{|i-j|}/(1-\rho^2)$와 같다. 따라서 $\mathbf{\Gamma}^{-1}$를 구체적으로 구하면 다음과 같다.

$$\mathbf{\Gamma}^{-1} = \begin{pmatrix} 1 & -\rho & 0 & \cdots & 0 \\ -\rho & 1+\rho^2 & -\rho & \cdots & 0 \\ 0 & -\rho & 1+\rho^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & -\rho & 1 \end{pmatrix}. \tag{2}$$

여기서 $\mathbf{\Gamma}^{-1}$는 $\boldsymbol{P}^T\boldsymbol{P}$로 표현되고, $\boldsymbol{P}$는 아래와 같다.

$$\boldsymbol{P} = \begin{pmatrix} \sqrt{1-\rho^2} & 0 & 0 & 0 & \cdots & 0 \\ -\rho & 1 & 0 & 0 & \cdots & 0 \\ 0 & -\rho & 1 & 0 & \cdots & 0 \\ 0 & 0 & -\rho & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix}. \tag{3}$$

따라서 $|\mathbf{\Gamma}|^{-1} = |\mathbf{\Gamma}^{-1}| = |\boldsymbol{P}|^2 = (1-\rho^2)$ 이며, $|\mathbf{\Gamma}|$는 $1/(1-\rho^2)$이다. 모형(1)에서, Rao와 Yu(1994)는 시계열 효과 요인 u_{it}에 대하여 AR(1) 모형을 적용하여 경험적 최량선형비편향추정량을 계산하였다. 여기서는 계층적 베이즈 추정량을 구하기 위하여 모형(1)을 아래와 같이 계층적 모형으로 변환하였다.

$$\begin{aligned} &\text{I. } \boldsymbol{y}_i|\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{v},r_1,r_2,\rho \overset{ind}{\sim} N(\boldsymbol{\theta}_i,\ \boldsymbol{\Sigma}_i) \\ &\text{II. } \boldsymbol{\theta}_i|\boldsymbol{\beta},\boldsymbol{v},r_1,r_2,\rho \overset{ind}{\sim} N(\boldsymbol{X}_i\boldsymbol{\beta}+v_i\mathbf{1}_T, {r_1}^{-1}\mathbf{\Gamma}) \\ &\text{III. } v_i|\boldsymbol{\beta},r_1,r_2,\rho \overset{ind}{\sim} N(0,{r_2}^{-1}) \\ &\text{IV. } \boldsymbol{\beta},\ \mathrm{r_1},\ \mathrm{r_2} \text{ and } \rho \text{ are mutually independent with} \\ &\qquad \boldsymbol{\beta} \sim \mathrm{Unif}(R^P) \\ &\qquad r_1 \sim \mathrm{Gamma}(\frac{1}{2}a_1,\frac{1}{2}b_1) \\ &\qquad r_2 \sim \mathrm{Gamma}(\frac{1}{2}a_2,\frac{1}{2}b_2) \\ &\qquad \rho \sim \mathrm{Unif}(-1,1), \end{aligned} \tag{4}$$

여기서 $\boldsymbol{\Sigma}_i$는 알려져 있으며, $r_1 = {\sigma_\eta}^{-2}$, $r_2 = \sigma_v^{-2}$이다. 확률변수 Z는 감마분포(α,β)를 따르고 있으며, 확률밀도함수는 $exp(-\alpha z)z^{\beta-1}$에 비례한다. (4)식으로 부터 $\boldsymbol{\theta}$, $\boldsymbol{v}$, $\boldsymbol{\beta}$, r_1, r_2,

ρ의 결합 사후확률밀도함수는

$$\begin{aligned} f(\boldsymbol{\theta}, \boldsymbol{v}, \boldsymbol{\beta}, r_1, r_2, \rho | \boldsymbol{y}) \propto & e^{-\frac{1}{2}\sum_{i=1}^{m}(\boldsymbol{y}_i-\boldsymbol{\theta}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i-\boldsymbol{\theta}_i)} \frac{{r_1}^{mT/2}}{|\boldsymbol{\Gamma}|^{m/2}} \\ & \times e^{-r_1\frac{1}{2}\sum_{i=1}^{m}(\boldsymbol{\theta}_i-\boldsymbol{X}_i\boldsymbol{\beta}-v_i\mathbf{1}_T)^T\boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta}_i-\boldsymbol{X}_i\boldsymbol{\beta}-v_i\mathbf{1}_T)} \\ & \times {r_2}^{m/2}e^{-r_2\sum_{i=1}^{m}{v_i}^2/2}e^{-a_1r_1/2}{r_1}^{b_1/2-1} \\ & \times e^{-a_2r_2/2} \times {r_2}^{b_2/2-1} \end{aligned} \tag{5}$$

와 같다. y_{ij} $(i = 1, \ldots, m; j = 1, \ldots, T)$가 주어졌을때, θ_{ij}의 사후확률분포를 계산하기 위해서, 고차원의 수치적분을 풀어야 하는 어려움이 있으나, 이같은 문제를 해결하기 위하여 깁스샘플러를 이용한다. 깁스샘플러를 이용하여 완전조건 사후확률분포들로부터 랜덤표본을 반복하여 생성하며, 생성된 샘플들로써 $\boldsymbol{\theta}$의 사후확률분포를 추정한다.

Gelman과 Rubin(1992)의 연구결과에 따라 깁스샘플러 체인을 n(≥2)개 실행하며, 각각은 2d번 반복하였으며, 각 체인별로 변수의 초기값은 과대산포분포(overdispersed distribution)로 부터 추출하였다. 초기 분포의 효과를 없애기 위해 각 체인의 처음 d번 반복은 분석에서 제외하였다.

깁스샘플러에 이용된 6개의 완전조건 사후확률분포들은 다음과 같다.

$$\begin{aligned} & \boldsymbol{\theta}_i|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{v}, r_1, r_2, \rho \\ & \overset{ind}{\sim} N\Big((\boldsymbol{\Sigma}_i^{-1} + r_1\boldsymbol{\Gamma}^{-1})^{-1}\{\boldsymbol{\Sigma}_i^{-1}\boldsymbol{y}_i + r_1\boldsymbol{\Gamma}^{-1}(\boldsymbol{X}_i\boldsymbol{\beta} + v_i\mathbf{1}_T)\}, \\ & \qquad\qquad (\boldsymbol{\Sigma}_i^{-1} + r_1\boldsymbol{\Gamma}^{-1})^{-1}\Big), \end{aligned} \tag{6}$$

$$\begin{aligned} & \boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{v}, r_1, r_2, \rho \\ & \sim N\Big((\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{\Gamma}^{-1}\boldsymbol{X}_i)^{-1}\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta}_i - v_i\mathbf{1}_T),\ (r_1\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{\Gamma}^{-1}\boldsymbol{X}_i)^{-1}\Big) \end{aligned} \tag{7}$$

$$\begin{aligned} & \boldsymbol{v}|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\beta}, r_1, r_2, \rho \\ & \overset{ind}{\sim} N\Big((r_1\mathbf{1}_T^T\boldsymbol{\Gamma}^{-1}\mathbf{1}_T + r_2)^{-1}r_1\mathbf{1}_T^T\boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta}_i - \boldsymbol{X}_i\boldsymbol{\beta}),\ \ (r_1\mathbf{1}_T^T\boldsymbol{\Gamma}^{-1}\mathbf{1}_T + r_2)^{-1}\Big) \end{aligned} \tag{8}$$

$$\begin{aligned} & r_1|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{v}, r_2, \rho \\ & \sim \text{Gamma}\Big(a_1/2 + \sum_{i=1}^{m}(\boldsymbol{\theta}_i - \boldsymbol{X}_i\boldsymbol{\beta} - v_i\mathbf{1}_T)^T\boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta}_i - \boldsymbol{X}_i\boldsymbol{\beta} - v_i\mathbf{1}_T), \\ & \qquad\qquad (mT + b_1)/2\Big) \end{aligned} \tag{9}$$

$$r_2|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{v}, r_1, \rho \sim \text{Gamma}\Big((a_2 + \sum_{i=1}^{m}v_i^2)/2, (m + b_2)/2\Big) \tag{10}$$

그리고

$$f(\rho|\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{v},\boldsymbol{\beta},r_1,r_2)$$
$$\propto |\boldsymbol{\Gamma}|^{-m/2}\mathrm{e}^{-r_1\sum_{i=1}^{m}(\boldsymbol{\theta}_i-\boldsymbol{X}_i\boldsymbol{\beta}-v_i\mathbf{1}_T)^T\boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta}_i-\boldsymbol{X}_i\boldsymbol{\beta}-v_i\mathbf{1}_T)}. \tag{11}$$

(6) ~ (10)의 완전조건 사후확률분포들은 알려진 분포이므로 직접적으로 표본을 생성할수 있으나, (11)의 분포로 부터 ρ를 생성하는 것은 그 분포를 알 수 없으므로 직접적으로 표본을 생성하는 것이 불가능하다. 그러므로 이러한 알려지지 않은 분포로 부터 ρ를 생성시키기 위하여 Metropolis-Hasting(M-H) 알고리즘을 이용하며, 결과적으로 "Metropolis-Hasting algorithm within Gibbs chains"의 형태로 표본을 추출하게 된다. 먼저 M-H 알고리즘을 실행하기 위해서는 적당한 candidate-generating density를 지정하여야 하는데, Chib와 Greenberg(1995)의 결과를 이용하여 candidate-generating density $q(\rho)$를 다음과 같이 구할 수 있다.

ρ의 사후확률분포를 $\pi(\rho)$라고 하자. 만약 $\pi(\rho)$ 가 두 함수의 곱의 형태로 나타내질 수 있다면, 즉, $\pi(\rho) \propto \psi(\rho)h(\rho)$, 여기에서 h($\rho$)는 알려진 확률밀도함수이며, $\psi(\rho)$는 균일하게 유계(uniformly bounded)된다고 할때, $q(\rho)$는 $h(\rho)$가 되며, probability of move는 $\alpha(x,y) = \min(\psi(y)/\psi(x),\ 1)$와 같다.

(11)식으로 부터 $\boldsymbol{\theta}_i - \boldsymbol{X}_i\boldsymbol{\beta} - v_i\mathbf{1}_T$의 j번째 요소 w_{ij}를 $\theta_{ij} - \boldsymbol{x}_{ij}^T\boldsymbol{\beta} - v_i$라고 두자. 그러면

$$\begin{aligned}&(\boldsymbol{\theta}_i - \boldsymbol{X}_i\boldsymbol{\beta} - v_i\mathbf{1}_T)^T\boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta}_i - \boldsymbol{X}_i\boldsymbol{\beta} - v_i\mathbf{1}_T)\\ &\quad = w_{i1}^2 + w_{iT}^2 + (1+\rho^2)(w_{i2}^2 + \cdots + w_{i,T-1}^2) - 2\rho\textstyle\sum_{j=1}^{T-1} w_{ij}w_{i,j+1}.\end{aligned}$$

여기서 $r_1 \times (w_{i1}^2 + w_{iT}^2) \equiv d_{i1}$, $r_1 \times (w_{i2}^2 + \cdots + w_{i,T-1}^2) \equiv d_{i2}$, 그리고 $r_1 \times \sum_{j=1}^{T-1} w_{ij}w_{i,j+1} \equiv d_{i3}$로 두면, (11)의 ρ의 사후확률밀도함수의 지수부분은 (12)식과 같이 계산된다.

$$\begin{aligned}&-r_1\sum_{i=1}^{m}(\boldsymbol{\theta}_i - \boldsymbol{X}_i\boldsymbol{\beta} - v_i\mathbf{1}_T)^T\boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta}_i - \boldsymbol{X}_i\boldsymbol{\beta} - v_i\mathbf{1}_T)\\ &\quad = \sum_{i=1}^{m}(-d_{i1} - d_{i2}(1+\rho^2) + 2\rho d_{i3})\\ &\quad = -d_1 - d_2(1+\rho^2) + 2\rho d_3.\end{aligned} \tag{12}$$

마지막 등식은 $\sum_{i=1}^{m} d_{i1} \equiv d_1$, $\sum_{i=1}^{m} d_{i2} \equiv d_2$, $\sum_{i=1}^{m} d_{i3} \equiv d_3$와 같이 둠으로써 성립된다. 그러므로 ρ의 사후확률밀도함수는 (13)식과 같이 나타난다.

$$f(\rho|\cdot) \propto (1-\rho^2)^{\frac{m}{2}}\exp[-d_2(\rho - \frac{d_3}{d_2})^2]. \tag{13}$$

그러므로, 여기에서 $q(\rho)$를 평균 d_3/d_2 과 분산 $1/(2d_2)$을 가지는 정규분포로 가져올 수 있으며, 함수 $\psi(\rho)$는 $(1-\rho^2)^{m/2}$가 되며, $\alpha(\rho^j, \rho^{j+1})$는 $\min((1-\rho^{j+1})^{m/2}/(1-\rho^j)^{m/2},\ 1)$이 된다. 여기에서 우리는 (11)의 분포형태에 대해서 관심을 가지게 되는데, 그 형태는 $\psi(\rho)$와 $q(\rho)$가 모두 오목(concave)한 형태이므로 ρ의 사후확률분포가 1개이상의 최빈값(mode)를 가질 것으로 예상된다.

깁스 샘플링의 결과로서 θ_{ij}의 사확률분포함수는 다음과 같이 추정된다.

$$\pi(\theta_{ij}|\boldsymbol{y}) \approx (nd)^{-1}\sum_{k=1}^{n}\sum_{l=d+1}^{2d}[\theta_{ij}|\boldsymbol{y},\boldsymbol{\beta}=\boldsymbol{\beta}_{kl},v_i=v_{ikl},r_1=r_{1kl},r_2=r_{2kl},\rho=\rho_{kl}]. \quad (14)$$

또한, Gelfand과 Smith(1991)의 "Rao-Blackwellized" 형태의 θ_{ij}의 사후확률분포의 평균과 분산은 다음과 같다.

$$\begin{aligned}E(\boldsymbol{\theta}_i|\boldsymbol{y}) \approx & (nd)^{-1}\sum_{k=1}^{n}\sum_{l=d+1}^{2d}(\boldsymbol{\Sigma}_i^{-1}+r_{1kl}\boldsymbol{\Gamma}^{-1})^{-1}\{\boldsymbol{\Sigma}_i^{-1}\boldsymbol{y}_i \\ & + r_{1kl}\boldsymbol{\Gamma}^{-1}(\boldsymbol{X}_i\boldsymbol{\beta}_{kl}+v_{ikl}\mathbf{1}_T)\}. \end{aligned} \quad (15)$$

$$\begin{aligned}V(\boldsymbol{\theta}_i|\boldsymbol{y}) \approx & (nd)^{-1}\sum_{k=1}^{n}\sum_{l=d+1}^{2d}(\boldsymbol{\Sigma}_i^{-1}+r_{1kl}\boldsymbol{\Gamma}^{-1})^{-1}+(\boldsymbol{\Sigma}_i^{-1}+r_{1kl}\boldsymbol{\Gamma}^{-1})^{-1} \\ & \times\{\boldsymbol{\Sigma}_i^{-1}\boldsymbol{y}_i+r_{1kl}\boldsymbol{\Gamma}^{-1}(\boldsymbol{X}_i\boldsymbol{\beta}_{kl}+v_{ikl}\mathbf{1}_T)\} \\ & \times\{\boldsymbol{\Sigma}_i^{-1}\boldsymbol{y}_i+r_{1kl}\boldsymbol{\Gamma}^{-1}(\boldsymbol{X}_i\boldsymbol{\beta}_{kl}\times v_{ikl}\mathbf{1}_T)\}^T \\ & \times(\boldsymbol{\Sigma}_i^{-1}+r_{1kl}\boldsymbol{\Gamma}^{-1})^{-1} \\ & -(nd)^{-2}\Bigg(\sum_{k=1}^{n}\sum_{l=d+1}^{2d}(\boldsymbol{\Sigma}_i^{-1}+r_{1kl}\boldsymbol{\Gamma}^{-1})^{-1} \\ & \qquad\times\{\boldsymbol{\Sigma}_i^{-1}\boldsymbol{y}_i+r_{1kl}\boldsymbol{\Gamma}^{-1}(\boldsymbol{X}_i\boldsymbol{\beta}_{kl}+v_{ikl}\mathbf{1}_T)\}\Bigg) \\ & \times\Bigg(\sum_{k=1}^{n}\sum_{l=d+1}^{2d}(\boldsymbol{\Sigma}_i^{-1}+r_{1kl}\boldsymbol{\Gamma}^{-1})^{-1} \\ & \qquad\times\{\boldsymbol{\Sigma}_i^{-1}\boldsymbol{y}_i+r_{1kl}\boldsymbol{\Gamma}^{-1}(\boldsymbol{X}_i\boldsymbol{\beta}_{kl}+v_{ikl}\mathbf{1}_T)\}\Bigg)^T. \end{aligned} \quad (16)$$

3. 모형 적합

이 절에서는 모형적합성 판정을 위해 Gelman 등 (1996)에 의해 소개된 posterior predictive assessment approach를 설명한다. 이 방법은 사후확률분포에서 생성된 샘플의 discrepancy measure와 관측 자료로부터 구한 discrepancy measure를 비교하며, 여기에 이용되는 discrepancy measure는 다음과 같다.

$$d(\boldsymbol{y},\boldsymbol{\theta}) = \sum_{i=1}^{m}(\boldsymbol{y}_i-\boldsymbol{\theta}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i-\boldsymbol{\theta}_i).$$

여기에서 $\boldsymbol{y}_{obs}$와 $\boldsymbol{y}_{new}$는 각각 관측값과 생성된 값을 가리키며, $f(\boldsymbol{\theta}|\boldsymbol{y}_{obs})$, $f(d(\boldsymbol{y}_{obs},\boldsymbol{\theta})|\boldsymbol{y}_{obs})$, $f(d(\boldsymbol{y}_{new},\boldsymbol{\theta})|\boldsymbol{y}_{obs})$는 $\boldsymbol{\theta}$, $d(\boldsymbol{y}_{obs},\boldsymbol{\theta})$ 와 $d(\boldsymbol{y}_{new},\boldsymbol{\theta})$의 사후확률분포이다. 실행과정은 다음과 같

다.

- Generate $\boldsymbol{\theta}^{(i)}$ from $f(\boldsymbol{\theta}|\boldsymbol{y}_{obs})$ and $\boldsymbol{y}^{(i)}$ from $f(\boldsymbol{y}|\boldsymbol{\theta}^{(i)})$
- Calculate the discrepancy measures $d(\boldsymbol{y}_{obs}, \boldsymbol{\theta}^{(i)})$ and $d(\boldsymbol{y}^{(i)}, \boldsymbol{\theta}^{(i)})$ for $i = 1, \ldots, B$, where B is the total number of Gibbs iterations of the $\boldsymbol{\theta}$ values.
- Approximate $P\{d(\boldsymbol{y}_{new}, \boldsymbol{\theta}) \geq d(\boldsymbol{y}_{obs}, \boldsymbol{\theta})|\boldsymbol{y}_{obs}\}$ by $B^{-1}\sum_{i=1}^{B} I\{d(\boldsymbol{y}^{(i)}, \boldsymbol{\theta}^{(i)}) \geq d(\boldsymbol{y}_{obs}, \boldsymbol{\theta}^{(i)})\}$ from the simulated values.

$P\{d(\boldsymbol{y}_{new}, \boldsymbol{\theta}) \geq d(\boldsymbol{y}_{obs}, \boldsymbol{\theta}|\boldsymbol{y}_{obs})\}$의 극값은 1과 0이며, 극값에 가까운 값이 나올경우 모형은 부적절한 것으로 판단하며, 반면 0.5에 가까울수록 모형은 적합하다.

4. 실제자료 분석

미국의 50개 주와 콜럼비아 자치지구에 대한 4인가족 중위소득 추정은 저소득 가구의 에너지 보조 정책 결정 등 연방정부의 다양한 정책결정에서 매우 중요한 통계이다.

기초자료는 매년 실시하는 CPS(Current Population Survey)자료로서, 자료의 형태는 조사가구내 15세이상 소득이있는 가구원의 소득을 모두 조사하여 가구별로 합산한 것으로 2,500달러 단위로 분류되어 있다. 그러므로 이를 선형보간법으로 보정하여 주별 4인가족 중위소득을 추정한다. 그러나, 주별로 할당되는 표본수가 적음으로, 추정치의 분산 또는 변동계수가 허용범위를 벗어나게 되어 직접적인 이용이 어려우므로 추정의 효율을 높이는 추정 방안이 요구된다. 주별 4인가족 중위소득의 추정을 위한 보조정보로는 먼저, 10년단위로 실시하는 총조사의 결과에서 얻어지는 중위소득과 경제분석국에서 매년 주별로 발표하는 1인당 소득이 있다. 이들 보조정정보는 전수조사 결과 또는 여러 경제지표들을 가공하여 작성한 것이므로 모두 표본오차는 생각 할 수 없다.

미국 센서스국에서는 1970년대 후반부터 주별 연간 4인가족소득의 중위값을 추정해오고 있으며, 센서스국에서 처음 이용한 방법은 10년간격 총조사에서 구해진 중위소득을 보완하여, 이값을 변량을 하는 회귀분석 모형을 통해 추정하는 것이었다. 10년단위로 조사되는 센서스에서 구해진 소득을 매년 경제분석국에서 발표하는 1인당 소득의 증감율을 적용하여 보완하였다. 현재, 센서스국에서는 Fay(1987)와 Fay 등(1993)의 경험적 베이즈 추정방법을 사용하고 있는데, 이 방법은 이변량 베이지안 모형으로서, 독립변수로는 CPS의 4인가족 중위소득의 자료와 3인과 5인 가족 소득의 선형결합 자료를 모두 이용하여, 이들 자료를 센서스자료 및 1인당 소득자료로서 보완하여 사용하고 있다. 그러나 이들 두 변량은 상관성이 높으므로 다중공선성의 문제를 안고 있어 이러한 문제의 해결방안으로 Ghosh, Nangia 그리고 Kim(1996)은 매 10년단위로 실시하는 4인가족 중위소득만을 보완하여 회귀식에 포함할것을 제안하였다.

이 절에서는 앞에서 제시된 계층적 베이즈 추정량으로 미국의 1989년의 주별 4인가족 소득 추정을 해보고자 하며, 이용된 자료는 9년간(1981 ~ 1989) CPS조사의 4인가족 소득의 중위값이다. 여기서 y_{it}는 i지역 t년의 4인 가족 소득 중앙값 θ_{it}의 CPS 추정량이다.

Ghosh 등(1996)의 권고에 따라 4인가족 소득의 중앙값만을 아래와 같이 보완하여 독립

변수로 포함하였다.

$$\text{조정된 센서스 중앙값} = \left[\frac{\text{BEA PCI}(c)}{\text{BEA PCI}(b)}\right] \times \text{census median}, \quad (17)$$

여기에서 BEA PCI(c)와 BEA PCI(b)는 경제분석국에서 작성하는 1인당 개인소득이며, 이때, c는 추정하고자 하는 연도를, b는 기준년을 뜻한다. 분석 모형은 2절에서 제시된 바와 같다.

$$y_{it} = \boldsymbol{x}_{it}^T\boldsymbol{\beta} + v_i + u_{it} + e_{it},\ i = 1,\cdots,m,\ j = 1,\cdots,T$$
$$u_{it} = \rho u_{i,t-1} + \eta_{it},\ |\rho| < 1.$$

깁스 샘플러를 이용, (6)~(11)의 완전조건사후확률분포로 부터 표본을 추출 하며, Gelman과 Rubin(1992)의 권고대로, 10개의 독립적인 깁스 표본자 체인으로부터 각각 4,000개를 샘플링하여, 이중 처음 구해진 앞의 2,000개는 버리고 나머지 2,000개만 분석에 이용하였다.

모형적합성 판정을 위해 깁스샘플링으로 부터 생성된 값과 실제 관측 자료값으로 부터 각각 discrepancy measure를 계산하였다. $B^{-1}\sum_{i=1}^{B} I\{d(\boldsymbol{y}^{(i)},\boldsymbol{\theta}^{(i)}) \geq d(\boldsymbol{y}_{obs},\boldsymbol{\theta}^{(i)})\}$의 계산값이 0.492로서 설정된 모형은 적절하다고 판단된다.

표1에서는, 이 논문에서 제시된 추정치와 기 제시된 추정치들을 비교하기 위하여 평균절대상대오차(c_1), 평균제곱상대오차(c_2), 평균절대오차(c_3)와 평균제곱편차(c_4)의 4개 기준을 이용하였다. 비교될 추정치들은 표본조사의 결과로 부터 직접 구해진 CPS 추정치, 센서스국에서 현재 사용하고 있는 경험적 베이즈 추정치, Ghosh 등(1996)에서 소개된 추정량(HB_1), 그리고 Datta 등(2002)의 경험적 베이즈 추정치이다. Datta 등(2002)은 알려지지 않은 초모수를 ML과 REML방법으로 추정하였다.

그 결과 본 논문에서 제시된 계층적 베이즈 추정량이 다른 추정량들보다 위 4개의 기준에서 모두 우세하였으며, 표본조사 결과로 부터 직접적으로 추정되는 CPS 추정치가 가장 좋지 못한 것으로 나타나 소지역추정모형 적용의 필요성을 잘 나타내주고 있다.

Table 1. 추정량의 효율성 비교

Estimate	c1	c2	c3	c4
Ours	0.0253	0.0011	1,023	1,968,334
CPS	0.0735	0.0084	2,929	13,811,122
Bureau	0.0296	0.0013	1,184	2,151,350
HB_1	0.0338	0.0018	1,352	3,095,736
EB(ML)	0.0278	0.0014	1,119	2,339,959
EB(REML)	0.0291	0.0014	1,126	2,368,397

참고문헌

[1] Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm, *The American Statistician*, 49, 327-335.

[2] Datta, G. S., Lahiri, P. and Maiti, T. (2002). Empirical Bayes Estimation of Median Income of Four-Person Families by State using Time Series and Cross-Sectional Datta, *Jornal of Statistical Planning and Inference*, 102, 83-97.

[3] Fay, R. E. (1987). "Application of Multivariate Regression to Small Domain Estimation", in *Small Area Statistics*, eds. R. Platek, J. N. K. Rao, C. E. Sarndal, and M. P. Singh, New York : Wiley, 91-102.

[4] Fay, R. E. and Herriot, R. A. (1979). Estimation of Income for Small Places: An application of James-Stein Procedures to Census data, *Journal of the American Statistical Association*, 74, 269-277.

[5] Fay, R. E., Nelson, C. T., and Litow, L. (1993), "Estimation of Median Income for 4-person Families by State", in *Indirect Estimators in Ferderal Programs*, Statistical Policy working Paper 21, Washing, D. C. : Statistical Policy Office, Office of Management and Budget, 901-917.

[6] Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities, *Journal of the American Statistical Association*, 85, 398-409.

[7] Gelfand, A. E. and Smith, A. F. M. (1991). Gibbs Sampling for Marginal Posterior Expectations, *Communications in Statistics-Theory and Method*, 20, 1747-1766.

[8] Gelman, A., Meng, X. -L., and Stern, H. (1996). Posterior Predictive Assessment of Model Fitness via Realized Discrepancies (with discussion), *Statistica Sinica*, 6, 733-807.

[9] Gelman, A. and Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences, *Statistical Science*, 7, 457-511.

[10] Ghosh, M., Nangia, N. and Kim, D. H. (1996). Estimation of Median Income of Four-Person Families: A Bayesian Time Series Approach, *Journal of the American Statistical Association*, 91, 1423-1431.

[11] Rao, J. N. K. and Yu, M. (1994). Small-area Estimation by Combining Time-series and Cross Sectional Data, *The Canadian Journal of Statistics*, 22, 511-528.

[2001년 10월 접수, 2002년 3월 채택]

Hierachical Bayes Estimation of Small Area Means in Repeated Survey

Dal-Ho Kim[1] Nam-hee Kim [2]

ABSTRACT

In this paper, we consider the HB estimators of small area means with repeated survey. Rao and Yu(1994) considered small area model with repeated survey data and proposed empirical best linear unbiased estimators. We propose a hierachical Bayes version of Rao and Yu by assigning prior distributions for unknown hyperparameters. We illustrate our HB estimator using very popular data in small area problem and then compare the results with the estimator of Census Bureau and other estimators previously proposed.

Keywords: Small area estimation; Hierachical Bayes model; Gibbs sampler; Reapeated survey.

1) Department of Statistics, Kyungpook National University, Taegu, 702-701, Korea
2) Planning Office, Province of Kyongbuk, Taegu, 702-702, Korea.