The Korean Journal of Applied Statistics (2012)
25*(3), 457–464*
DOI: http://dx.doi.org/10.5351/KJAS.2012.25.3.457

Small Area Estimation via Nonparametric Mixed Effects Model

Seok-Oh Jeong[1] · Key-Il Shin[2]

[1]Department of Statistics, Hankuk University of Foreign Studies
[2]Department of Statistics, Hankuk University of Foreign Studies

(Received April 30, 2012; Revised May 17, 2012; Accepted June 4, 2012)

Abstract

Small area estimation is a statistical inference method to overcome the large variance due to the small sample size allocated in a small area. Recently some nonparametric estimators have been applied to small area estimation. In this study, we suggest a nonparametric mixed effect small area estimator using kernel smoothing and compare the small area estimators using labor statistics.

Keywords: Small area estimation, mixed effects model, nonparametric mixed effects model, kernel smoothing.

1. 서론

최근 관심이 집중되는 통계 분야의 하나로 소지역추정(small area estimation)이 부상하고 있다. 소지역추정을 간단히 소개하면 지역 또는 도메인에 배분된 표본의 수가 작아 정확한 소지역추정이 불가능할 때 이를 극복하는 통계적 방법이다. 우리나라 뿐만 아니라 세계적으로 작은 지역 또는 도메인에 관한 통계를 정확히 구하려는 움직임이 있고 이를 뒷받침하기 위한 통계적 기법들이 개발되고 있다.

소지역추정법은 크게 자료기반(data-based) 또는 설계기반(design-based) 추정법과 모형기반(model-based) 추정법으로 나누어진다. 설계기반 추정법은 주어진 자료만을 사용하고 추가적인 정보를 사용하지 않기 때문에 추정의 정확도를 향상시키는 데 한계가 있어 최근의 분석에서는 잘 사용하지 않는 방법이다. 반면에 모형기반 추정법은 추가적인 정보, 즉 보조정보(auxiliary information)를 이용하기 때문에 보다 정밀한 추정이 가능하다. 모형기반 추정법으로는 기본적으로 회귀추정법, 비추정법과 같은 일반화회귀추정법(generalized regression method)이 사용된다. 최근 개발된 고급 통계 기법으로는 선형혼합모형 추정법, 계층적베이즈 추정법 등이 있다. 이상의 방법들은 모두 모수적 모형에 의한 것으로 이미 많은 소지역추정에서 사용되고 있다.

모형기반 추정법에서는 보조 정보의 양이 클수록 정확한 소지역추정 방법이 가능하므로 보조 정보의 양을 증가시키는 방법이 최근 연구되었다. 공간 정보를 이용하여 소지역추정법의 정확도를 향상시키는 방

This research was supported by the Basic Science Research Program of the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (2010-0008807).
[2]Corresponding author: Department of Statistics, Hankuk University of Foreign Studies, Yongin, Gyeonggi, 449-791, Korea. E-mail: keyshin@hufs.ac.kr

법이 Kim 등 (2008)에 의해 연구되었으며, 단위수준정보(unit level information)와 지역수준정보(area level information)를 결합하여 추정의 정도를 높이는 방법이 Lee와 Shin (2012)에서 연구되었다. 물론 시계열 분석 모형을 이용한 방법 또한 연구되고 있다. 이에 관한 내용은 Kim 등 (2005)을 살펴보기 바란다.

많은 학자들은 모수적 모형은 이미 완성 단계 혹은 한계에 이르렀다고 생각하고 있어 모수적 모형을 넘어선 다른 방법을 통해 모형의 정확도를 향상시키려는 움직임을 보이고 있는데 비모수적 함수 추정을 이용한 방법이 유망할 것으로 전망된다. 최근 들어 비모수적 방법이 소지역추정법에 적용되기 시작하였는데, Opsomer 등 (2008), Salvati 등 (2010)이 비모수적 방법을 이용한 소지역추정에 관하여 연구 결과를 발표하였다. 이들은 선형혼합모형을 스플라인 평활을 이용해 비모수적 방법으로 확장한 것이다. 본 논문에서는 커널 평활을 이용해 기존의 선형혼합모형에 기초한 모형기반 소지역추정법을 비모수적 모형으로 확장하는 방법을 제시하고 그 유효성을 실증하고자 한다.

본 논문은 다음과 같이 구성되었다. 2절에서 선형혼합모형을 이용한 모형기반 소지역추정법을 간단히 설명하고, 모형의 고정효과 부분을 커널 평활을 이용해 비모수적 모형으로 확장하는 방법을 제안한다. 3절에서 모의실험을 통하여 제안된 추정량의 우수성을 확인하였으며, 4절에서 결론을 맺는다.

2. 비모수혼합모형을 이용한 소지역추정

본 논문에서는 소지역추정에서 사용되는 지역수준자료(area level data)와 단위수준자료(unit level data) 중에서 단위수준자료에 관하여 연구하였다. 또한 전 논문을 통하여 다음과 같은 설계를 사용하였다. 크기가 N인 모집단 U가 d개의 소지역의 모집단 U_j, $j = 1, 2, \ldots, d$으로 구성되어 있다고 하자. j번째 소지역의 모집단 U_j의 크기를 N_j라 하면 $\sum_{j=1}^{d} N_j = N$가 된다. 연속형인 관심변수 y에 대해 크기 n인 표본을 추출하되 각 소지역에서 얻어진 표본 크기를 n_j라 하면, $\sum_{j=1}^{d} n_j = n$가 된다. 또한 s_j를 j번째 소지역에서 추출된 표본 집합, r_j를 이 소지역에서 표본조사에서 제외된 집합이라 하면, $U_j = s_j \cup r_j$, $j = 1, 2, \ldots, d$이다. 본 논문의 연구 목적은 표본집합 s_j의 관심변수와 보조정보만을 이용하여 각 소지역 U_j에서 관심변수 y의 평균을 추정하는 것이다.

2.1. 선형혼합모형(linear mixed effect model)

j번째 소지역에서 i번째 관측치를 y_{ij}라 할 때 일반적으로 사용되고 있는 선형혼합모형은 다음과 같다.

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{z}_{ij}^T \boldsymbol{\gamma}_j + \epsilon_{ij}, \quad i = 1, 2, \ldots, n_j, \ j = 1, 2, \ldots, d \tag{2.1}$$

여기서 $\mathbf{x}_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijp})^T$와 $\mathbf{z}_{ij} = (z_{ij1}, z_{ij2}, \ldots, z_{ijp})^T$는 보조변수 벡터, $\boldsymbol{\beta}$는 고정효과(fixed effects), $\boldsymbol{\gamma}_j \sim N(\mathbf{0}_q, \mathbf{G})$는 지역에 따른 랜덤효과(area-specific random effect), $\epsilon_i \sim N(0, \sigma_\epsilon^2)$는 오차항이다. $\mathbf{0}_m$은 모든 성분이 0이고 길이가 m인 벡터를, $\mathbf{1}_m$은 모든 성분이 1이고 길이가 m인 벡터를 나타낸다. 이를 각 소지역별로 묶어 행렬 및 벡터 기호로 나타내면 다음과 같다.

$$\mathbf{y}_j = \mathbf{X}_j \boldsymbol{\beta} + \mathbf{Z}_j \boldsymbol{\gamma}_j + \boldsymbol{\epsilon}_j, \quad j = 1, 2, \ldots, d \tag{2.2}$$

$$\mathbf{y}_j = \begin{pmatrix} y_{1j} \\ y_{2j} \\ \vdots \\ y_{n_j j} \end{pmatrix}, \quad \mathbf{X}_j = \begin{pmatrix} \mathbf{x}_{1j}^T \\ \mathbf{x}_{2j}^T \\ \vdots \\ \mathbf{x}_{n_j j}^T \end{pmatrix}, \quad \mathbf{Z}_j = \begin{pmatrix} \mathbf{z}_{1j}^T \\ \mathbf{z}_{2j}^T \\ \vdots \\ \mathbf{z}_{n_j j}^T \end{pmatrix}, \quad \boldsymbol{\epsilon}_j = \begin{pmatrix} \epsilon_{1j} \\ \epsilon_{2j} \\ \vdots \\ \epsilon_{n_j j} \end{pmatrix} \sim N(\mathbf{0}_{n_j}, \mathbf{R}_j).$$

이들을 다시 각 소지역에 대해 열방향으로 쌓아올려 단번에 나타내면

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon} \tag{2.3}$$

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_d \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_d \end{pmatrix}, \quad \mathbf{Z} = \mathrm{diag}(\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_d),$$

$$\boldsymbol{\gamma} = \begin{pmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \\ \vdots \\ \boldsymbol{\gamma}_d \end{pmatrix} \sim N\left(\mathbf{0}_{qd}, \bar{\mathbf{G}}\right), \quad \boldsymbol{\epsilon} = \begin{pmatrix} \boldsymbol{\epsilon_1} \\ \boldsymbol{\epsilon_2} \\ \vdots \\ \boldsymbol{\epsilon_d} \end{pmatrix} \sim N(\mathbf{0}_N, \mathbf{R})$$

이 되는데 서로 다른 소지역 간의 랜덤효과 및 오차항이 서로 독립임을 가정하면 $\overline{\boldsymbol{G}} = \mathrm{diag}(\mathbf{G}, \mathbf{G}, \ldots, \mathbf{G})$, $\mathbf{R} = \mathrm{diag}(\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_d)$가 된다.

랜덤효과 $\boldsymbol{\gamma}$와 오차항 $\boldsymbol{\epsilon}$이 서로 독립임을 가정하자. 각 분산성분 $\mathbf{G}$과 $\mathbf{R}_j$이 주어진 경우 로그우도함수는

$$\begin{aligned} & l(\boldsymbol{\beta}, \boldsymbol{\gamma} | \mathbf{y}, \mathbf{X}) \\ & = -\frac{1}{2} \sum_{j=1}^{d} \left\{ (\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta} - \mathbf{Z}_j\boldsymbol{\gamma_j})^T \mathbf{R}_j^{-1} (\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta} - \mathbf{Z}_j\boldsymbol{\gamma}_j) + \boldsymbol{\gamma}_j \mathbf{G}^{-1} \boldsymbol{\gamma}_j + \log|\mathbf{G}| + \log|\mathbf{R}_j| \right\} \end{aligned} \tag{2.4}$$

와 같이 주어지게 되어 이를 최대로 하는 $(\boldsymbol{\beta}, \boldsymbol{\gamma})$를 구하면

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y},$$
$$\hat{\boldsymbol{\gamma}}_j = \mathbf{G}\mathbf{Z}_j^T \mathbf{V}_j^{-1} \left(\mathbf{y}_j - \mathbf{X}_j \hat{\boldsymbol{\beta}}\right)$$

와 같다. 단, $\mathbf{V}_j = \mathbf{Z}_j \mathbf{G} \mathbf{Z}_j^T + \mathbf{R}_j$, $\mathbf{V} = \mathrm{diag}(\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_d) = \mathbf{Z}\overline{\mathbf{G}}\mathbf{Z}^T + \mathbf{R}$이다. 이 추정과정에서 필요한 분산공분산행렬 R과 G는 최대우도추정법(ML) 또는 제한적 최대우도추정법(restricted ML, ReML) 등을 이용하여 얻을 수 있다.

이를 이용하면 조사되지 않은 관심 변수 y_{ij}, $i \in r_j$의 예측치는 $\mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} + \mathbf{z}_{ij}^T \hat{\boldsymbol{\gamma}}_j$와 같이 구할 수 있고, 따라서 지역별 평균 $\overline{Y}_j$, $j = 1, 2, \ldots, d$의 소지역추정량은 다음과 같다.

$$\hat{\overline{Y}}_j^{MX} = N_j^{-1} \left\{ \sum_{i \in s_j} y_{ij} + \sum_{i \in r_j} \left(\mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} + \mathbf{z}_{ij}^T \hat{\boldsymbol{\gamma}}_j \right) \right\}, \quad j = 1, 2, \ldots, d. \tag{2.5}$$

이상에 관한 자세한 내용은 Rao (2003)을 참조하기 바란다.

2.2. 비모수혼합모형(nonparametric mixed effects model)

이 절에서는 다음과 같은 비모수혼합모형에 기반한 소지역추정법을 논하고자 한다.

$$y_{ij} = \eta(x_{ij}) + \gamma(x_{ij}) + \epsilon_{ij}, \quad i = 1, 2, \ldots, n_j, \; j = 1, 2, \ldots, d \tag{2.6}$$

여기서 $\eta(\cdot)$는 고정효과를 나타내는 함수, $\gamma_j(\cdot)$는 소지역 j에 해당하는 랜덤효과를 나타내는 함수이다. 문제를 단순화하기 위하여 보조변수들은 단변량인 경우로 한정해 논의를 진행한다. 함수 η와 γ_j가 매끈하면 $x_{ij} \approx x$일 때

$$\eta(x_{ij}) \approx \eta(x) + \eta'(x)(x_{ij}-x) + \cdots + \frac{\eta^{(p)}(x)}{p!}(x_{ij}-x)^p = \widetilde{\mathbf{x}}_{ij}^T\widetilde{\boldsymbol{\beta}},$$

$$\gamma(x_{ij}) \approx \gamma(x) + \gamma'(x)(x_{ij}-x) + \cdots + \frac{\gamma^{(q)}(x)}{q!}(x_{ij}-x)^q = \widetilde{\mathbf{z}}_{ij}^T\widetilde{\boldsymbol{\gamma}},$$

$$\widetilde{\mathbf{x}}_{ij} = \begin{pmatrix} 1 \\ x_{ij}-x \\ \vdots \\ (x_{ij}-x)^p \end{pmatrix}, \quad \widetilde{\mathbf{z}}_{ij} = \begin{pmatrix} 1 \\ x_{ij}-x \\ \vdots \\ (x_{ij}-x)^q \end{pmatrix}, \quad \widetilde{\boldsymbol{\beta}} = \begin{pmatrix} \eta(x) \\ \eta'(x) \\ \vdots \\ \eta^{(p)}(x)/p! \end{pmatrix}, \quad \widetilde{\boldsymbol{\gamma}}_j = \begin{pmatrix} \gamma(x) \\ \gamma'(x) \\ \vdots \\ \gamma^{(q)}(x)/q! \end{pmatrix}$$

와 같은 근사가 성립한다. 결국 비모수혼합모형과 선형혼합모형을 비교하면 고정효과 부분과 랜덤효과 부분에 포함된 행렬과 벡터의 원소가 다를 뿐 두 수식의 형태가 일치한다는 것을 확인할 수 있다. 따라서 식 (2.6)의 비모수적혼합모형은 다음과 같은 선형혼합모형으로 근사가 가능하다. 즉 $x_{ij} \approx x$일 때

$$y_{ij} = \widetilde{\mathbf{x}}_{ij}^T\widetilde{\boldsymbol{\beta}} + \widetilde{\mathbf{z}}_{ij}^T\widetilde{\boldsymbol{\gamma}}_{\boldsymbol{j}} + \epsilon_{ij}, \tag{2.7}$$

$$\mathbf{y}_j = \widetilde{\mathbf{X}}_j^T\widetilde{\boldsymbol{\beta}} + \widetilde{\mathbf{Z}}_j^T\widetilde{\boldsymbol{\gamma}}_{\boldsymbol{j}} + \boldsymbol{\epsilon}_j, \tag{2.8}$$

$$\mathbf{y} = \widetilde{\mathbf{X}}^T\widetilde{\boldsymbol{\beta}} + \widetilde{\mathbf{Z}}^T\widetilde{\boldsymbol{\gamma}} + \boldsymbol{\epsilon}, \tag{2.9}$$

$$\widetilde{\mathbf{y}}_j = \begin{pmatrix} y_{1j} \\ y_{2j} \\ \vdots \\ y_{n_jj} \end{pmatrix}, \quad \widetilde{\mathbf{X}}_j = \begin{pmatrix} \widetilde{\mathbf{x}}_{1j}^T \\ \widetilde{\mathbf{x}}_{2j}^T \\ \vdots \\ \widetilde{\mathbf{x}}_{n_jj}^T \end{pmatrix}, \quad \widetilde{\mathbf{Z}}_j = \begin{pmatrix} \mathbf{Z}_{1j}^T \\ \mathbf{Z}_{2j}^T \\ \vdots \\ \mathbf{Z}_{n_jj}^T \end{pmatrix}, \quad \widetilde{\boldsymbol{\epsilon}}_j = \begin{pmatrix} \boldsymbol{\epsilon}_{1j}^T \\ \boldsymbol{\epsilon}_{2j}^T \\ \vdots \\ \boldsymbol{\epsilon}_{n_jj}^T \end{pmatrix},$$

$$\widetilde{\mathbf{y}} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_d \end{pmatrix}, \quad \widetilde{\mathbf{X}} = \begin{pmatrix} \widetilde{\mathbf{X}}_1 \\ \widetilde{\mathbf{X}}_2 \\ \vdots \\ \widetilde{\mathbf{X}}_d \end{pmatrix}, \quad \widetilde{\mathbf{Z}} = \operatorname{diag}\left(\widetilde{\mathbf{Z}}_1, \widetilde{\mathbf{Z}}_2, \ldots, \widetilde{\mathbf{Z}}_d\right), \quad \widetilde{\boldsymbol{\gamma}} = \begin{pmatrix} \widetilde{\boldsymbol{\gamma}}_1 \\ \widetilde{\boldsymbol{\gamma}}_2 \\ \vdots \\ \widetilde{\boldsymbol{\gamma}}_d \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_d \end{pmatrix}.$$

이제 식 (2.6)의 비모수적혼합모형을 식 (2.1)에서 식 (2.3)의 선형혼합모형과 똑같은 형태로 근사시킬 수 있음을 알게 되었다. 추가로 $\widetilde{\boldsymbol{\gamma}}_j \sim N(\mathbf{0}_{q+1}, \widetilde{\mathbf{G}})$을 가정하고 나면 식 (2.4)의 선형혼합모형을 위한 로그우도함수를 다음과 같이 x-근방에서의 국소로그우도함수(local log-likelihood)로 확장해 사용할 수 있게 된다.

$$\begin{aligned} l_{x,h}\left(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\gamma}} | \mathbf{y}, \widetilde{\mathbf{X}}\right) = -\frac{1}{2}\sum_{j=1}^{d}\Big\{ &\left(\widetilde{\mathbf{y}}_j - \widetilde{\mathbf{X}}_j\widetilde{\boldsymbol{\beta}} - \widetilde{\mathbf{Z}}_j\widetilde{\boldsymbol{\gamma}}_j\right)^T \mathbf{K}_{j,h}^{\frac{1}{2}}\mathbf{R}_j^{-1}\mathbf{K}_{j,h}^{\frac{1}{2}}\left(\widetilde{\mathbf{y}}_j - \widetilde{\mathbf{X}}_j\widetilde{\boldsymbol{\beta}} - \widetilde{\mathbf{Z}}_j\widetilde{\boldsymbol{\gamma}}_j\right) \\ &+ \widetilde{\boldsymbol{\gamma}}_j^T\widetilde{\mathbf{G}}^{-1}\widetilde{\boldsymbol{\gamma}}_j + \log\left|\widetilde{\mathbf{G}}\right| + \log\left|\widetilde{\mathbf{R}}_j\right|\Big\}, \end{aligned} \tag{2.10}$$

$$\mathbf{K}_{j,h} = \operatorname{diag}\left\{K_h(x_{ij}-x),\, i = 1, 2, \ldots, n_j\right\}, \quad K_h(u) = h^{-1}K\left(\frac{u}{h}\right).$$

단, 커널(kernel) K는 0을 중심으로 대칭인 확률밀도함수, h는 평활모수(smoothing parameter)이다.

식 (2.10)을 최적화하여 얻어지는 $\widetilde{\boldsymbol{\beta}}$와 $\widetilde{\boldsymbol{\gamma}}_j$의 첫 성분의 값들이 각각 $\eta(x)$와 $\gamma_j(x)$의 추정치(혹은 예측치) $\widetilde{\eta}(x)$, $\widetilde{\gamma}_j(x)$가 되며 x에서의 Y의 예측치는 이들의 합으로 정하면 된다. 즉, 본 논문이 제안하는 지역별 평균 $\overline{Y}_j$, $j = 1, 2, \ldots, d$의 비모수적 소지역추정량은

$$\hat{\overline{Y}}_j^{NPMX} = N_j^{-1} \left\{ \sum_{i \in s_j} y_{ij} + \sum_{i \in r_j} (\hat{\eta}(x_{ij}) + \hat{\gamma}_j(x_{ij})) \right\}, \quad j = 1, 2, \ldots, d \tag{2.11}$$

와 같다.

위 근사식의 차수 p와 q를 높여 근사의 정밀도를 향상시킬 수 있지만 그에 따라 증대되는 복잡성에 비해 실질적으로 얻을 수 있는 이득이 거의 없을 것으로 판단되므로 다음 절의 모의실험 및 자료 분석에서는 고정효과에 대해서는 국소선형($p = 1$), 랜덤효과에 대해서는 국소상수($q = 0$) 모형을 고려한다. 또한 커널 K에 표준정규분포 $N(0, 1)$의 밀도함수인 $K(u) = (2\pi)^{-1/2} e^{-u^2/2}$을 사용했다.

한편 식 (2.10)의 국소로그우도를 최적화해 원하는 추정치 혹은 예측치를 얻는 문제는

$$y_{ij}^* = K_h(x_{ij} - x)^{\frac{1}{2}} y_j, \quad \mathbf{x}_{ij}^* = K_h(x_{ij} - x)^{\frac{1}{2}} \begin{pmatrix} 1 \\ x_{ij} - x \\ \vdots \\ (x_{ij} - x)^p \end{pmatrix}, \tag{2.12}$$

$$\mathbf{z}_{ij}^* = K_h(x_{ij} - x)^{\frac{1}{2}} \begin{pmatrix} 1 \\ x_{ij} - x \\ \vdots \\ (x_{ij} - x)^q \end{pmatrix}, \quad i = 1, 2, \ldots, n_j, \ j = 1, 2, \ldots, d$$

와 같이 자료값을 변환한 후 선형혼합모형을 적합시키는 것과 같은 문제임을 확인할 수 있다. 따라서 식 (2.12)와 같이 변환된 자료에 SAS의 PROC MIXED나 R의 lme( )와 같은 기존의 통계 계산 프로그램을 적용하면 된다 (Wu와 Zhang, 2006).

3. 2006 매월노동통계 자료를 이용한 모의실험

이 절에서는 본 논문에서 제안하는 비모수적혼합모형 소지역추정량의 성능을 기존의 선형혼합모형 소지역추정량과 비교하고자 한다. 분석에 사용한 자료는 노동부의 '2006년 매월노동통계'의 원자료이다. 이 자료는 전국의 $n = 7{,}038$ 사업체를 조사해 얻은 임금 총액(y) 및 종사자 수(x)의 자료이며, 소지역은 전국의 $d = 47$개 지청이 된다. 따라서 고정효과에 사용된 설명변수는 종사자 수이고 또한 랜덤효과인 지역효과를 반영하기 위하여 종사자 수를 각 소지역의 지역변수로 고려하였다. 각 소지역 내 표본 사업체 수를 n_j라 하면 $n = \sum_{j=1}^{d} n_j$가 된다. 각 사업체의 종사자 수를 보조변수로 하여 소지역별 평균 임금을 추정하는 상황을 가정하고, 다음의 절차에 따라 모의실험을 실시했다.

(1) 크기가 7,038인 원자료 중 종사자 수가 300 미만인 6,301개에 대해 5회의 재추출(복원 허용)을 실시한 후 종사자 수가 300명 이상인 737개의 사업체를 합쳐 크기가 $N = 32{,}242\,(= 6{,}3301 * 5 + 737)$인 의사모집단(pseudo population) U를 생성한다. 생성된 의사모집단을 각 소지역별로 구별하여 U_j라 하고 그 크기 N_j를 구한다.

(2) 생성된 의사모집단의 소지역별 평균 $\overline{Y}_j$, $j = 1, 2, \ldots, d$를 계산한다.

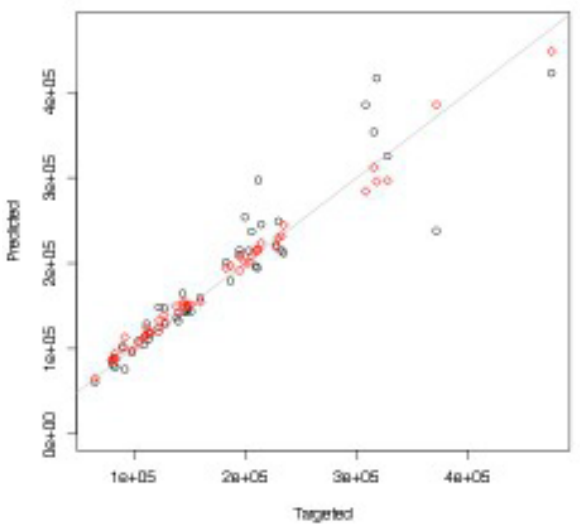

Figure 3.1. Mean predicted values $\hat{\overline{Y}}_j$ and target values $\overline{Y}_j(h{=}0.15)$ (Circles ($\overline{Y}_j$, $\hat{\overline{Y}}_j^{MX}$), Diamonds ($\overline{Y}_j$, $\hat{\overline{Y}}_j^{NPMX}$), Solid line $y = x$)

(3) 의사모집단 U로부터 각 소지역을 층으로 하는 층화추출을 통해 원자료와 크기가 $n = 7{,}038$로 동일한 표본을 얻는다. 단 각 표본에는 종사자 수가 300명 이상인 사업체 737개가 반드시 포함되도록 하며, 층화추출된 표본 내 각 층의 크기가 원자료와 동일하게 n_j가 되도록 한다. 각 층에서 추출된 자료를 모은 것을 s_j라 하면 $r_j = U_j - s_j$이 된다.

(4) 비교 대상인 소지역추정방법에 따라 (4)에서 추출된 표본으로 각 소지역별(지청별) 평균 임금 추정치 $\hat{\overline{Y}}_j$ 들을 구한 후 (3)의 $\overline{Y}_j$와 비교한다.

(5) (3)∼(4)를 $R = 500$회 반복 실시한다.

비교를 위한 통계량으로는 Rao (2003)에서 이용하고 있는 여러 비교통계량들을 사용했다. 즉, 오차의 크기에 근거한 Mean Squared Error(MSE)와 Mean Absolute Error(MAE), 상대적인 오차의 크기를 비교하기 위한 것으로 Relative Error(RE)와 Absolute Relative Error(ARE)를 고려한다. 각 비교통계량의 구체적 형태는 다음과 같다.

$$\mathrm{MSE} = \frac{1}{R}\frac{1}{d}\sum_{r=1}^{R}\sum_{j=1}^{d}\left(\hat{\overline{Y}}_j^{(r)} - \overline{Y}_j\right)^2,$$

$$\mathrm{MAE} = \frac{1}{R}\frac{1}{d}\sum_{r=1}^{R}\sum_{j=1}^{d}\left|\hat{\overline{Y}}_j^{(r)} - \overline{Y}_j\right|,$$

$$\mathrm{RE} = \frac{1}{R}\frac{1}{d}\sum_{r=1}^{R}\sum_{j=1}^{d}\left(\frac{\hat{\overline{Y}}_j^{(r)} - \overline{Y}_j}{\overline{Y}_j}\right)^2,$$

$$\mathrm{ARE} = \frac{1}{R}\frac{1}{d}\sum_{r=1}^{R}\sum_{j=1}^{d}\left|\frac{\hat{\overline{Y}}_j^{(r)} - \overline{Y}_j}{\overline{Y}_j}\right|,$$

Table 3.1. Comparison of the small area estimators using labor statistics

추정량	h	MSE($\times 10^9$)	MAE($\times 10^4$)	RE($\times 10^{-2}$)	ARE($\times 10^{-1}$)
선형혼합 $\hat{\overline{Y}}_j^{MX}$	∞	1.3022	2.1199	1.9828	1.0163
	(s.e)	(0.0004)	(0.0004)	(0.0014)	(0.0000)
비모수혼합 $\hat{\overline{Y}}_j^{NPMX}$	0.10	0.1189	0.8051	0.5251	0.5200
	(s.e)	(0.0006)	(0.0024)	(0.0030)	(0.0017)
	0.15	0.1178	0.8168	0.5462	0.5342
	(s.e)	(0.0006)	(0.0025)	(0.0035)	(0.0019)
	0.20	0.1191	0.8429	0.5830	0.5595
	(s.e)	(0.0006)	(0.0026)	(0.0036)	(0.0019)
	0.25	0.1220	0.8757	0.6291	0.5895
	(s.e)	(0.0007)	(0.0027)	(0.0058)	(0.0020)
	0.30	0.1299	0.9283	0.6900	0.6343
	(s.e)	(0.0007)	(0.0028)	(0.0058)	(0.0021)

단, $\hat{\overline{Y}}_j^{(r)}$ 은 r번째($r = 1, 2, \ldots, R$) 모의실험에서 얻은 소지역추정량을 뜻한다.

아래 Figure 3.1은 500회의 반복 실험 중 한 실험 결과를 도시한 것이다. 동그라미로 표시된 점들은 선형혼합모형에 의해 얻은 각 소지역 평균의 추정치 $\hat{\overline{Y}}_j^{MX}$ 와 소지역 평균 $\overline{Y}_j$의 순서쌍을, 마름모로 표시된 점들은 비모수혼합모형에 의해 얻은 소지역 평균 추정치 $\hat{\overline{Y}}_j^{NPMX}$ 와 소지역 평균 $\overline{Y}_j$의 순서쌍을 나타낸 것이고, 대각의 실선은 $y = x$의 직선으로 점들이 이 직선 근처에 있을수록 추정이 잘 된 것임을 의미한다. 그림에서 보는 바대로 본 논문이 제안하는 비모수혼합모형 소지역추정량의 성능이 선형혼합모형 소지역추정량에 비해 상당히 우수함을 알 수 있다. 물론 나머지 499개의 반복 실험 결과에서도 비슷한 패턴을 유지하였음을 확인했다.

다음의 Table 3.1은 500회의 반복실험을 통해 얻은 각종 비교통계량의 값들을 정리한 것으로, 이 표에서도 역시 비모수혼합모형에 의한 소지역추정량의 우수성을 확인할 수 있다. 예를 들어 MSE의 경우, 상당히 넓은 범위의 h값에 대해 제안된 방법이 기존의 선형혼합모형추정법에 비해 10배 이상 우수한 것을 확인할 수 있다.

4. 결론 및 전망

소지역추정 방법으로 모형기반 추정법에 대한 관심이 증대되는 가운데 비모수적 함수 추정 기법은 이 분야에서 활용도가 매우 높을 것으로 기대된다. 본 논문이 제안하는 국소다항모형을 이용한 비모수적혼합모형 소지역추정법은 2006 매월노동통계 자료에 대해 적용한 결과 매우 우수한 성능을 보였다. 비모수적 방법 고유의 유연성을 고려할 때 여타 자료에서도 여전히 우수한 성능을 보일 것으로 기대된다. 다만 본 연구의 모의실험에서 절사표본설계 방법이 사용되었다. 이는 자료의 극단 부분에 포함된 층의 자료수가 매우 적을 때 비모수적 함수추정을 이용하여 소지역추정을 실시할 경우 추정의 정도를 향상시기키 위한 것이다. 따라서 극단층의 표본크기가 추정의 정도에 미치는 영향력에 관한 추가적인 연구가 필요하다. 또한 일반적으로 비모수적함수추정 기법을 적용할 때 늘 그렇듯이 비모수적 혼합모형 적합 시 평활 모수(smoothing parameter)를 적절히 선택하는 것이 매우 중요한데, 본 논문에서는 심도있게 다루지 못했으며 이를 향후 연구 과제로 남겨둔다.

References

Kim, J.-D., Shin, K.-I. and Lee, S. E. (2005). Space time autoregressive model for small area estimation, *The Korean Journal of Applied Statistics*, **18**, 627–637.

Kim, J.-S., Hwang, H.-J. and Shin, K.-I. (2008). Comparison of spatial small area estimators based on neighborhood information systems, *The Korean Journal of Applied Statistics*, **21**, 855–866.

Lee, S. E. and Shin, K.-I. (2012). Two stage small area estimation, *The Korean Journal of Applied Statistics*, **25**, 293–300.

Opsomer, J. D., Claeskens, G., Ranalli, M. G., Kauermann, G. and Breidt, F. J. (2008). Non-parametric small area estimation using penalized spline regression, *Journal of Royal Statistical Society B*, **70**, 265–286.

Rao, J. N. K. (2003). *Small Area Estimation*, John Wiley & Sons.

Salvati, N., Chandra, H., Ranalli, M. G. and Chambers, R. (2010). Small area estimation using a nonparametric model-based direct estimator, *Computational Statistics and Data Analysis*, **54**, 2159–2171.

Wu, H. and Zhang, J.-T.(2006). *Nonparametric Regression Methods for Longitudinal Data Analysis*, John Wiley & Sons.