*The Korean Journal of Applied Statistics* (2013) DOI: http://dx.doi.org/10.5351/KJAS.2013.26.1.071
**26**(1), 71–79

# Semiparametric and Nonparametric Mixed Effects Models for Small Area Estimation

Seok-Oh Jeong[a] · Key-Il Shin[a,1]

[a]Department of statistics, Hankuk University of Foreign Studies

(Received October 12, 2012; Revised December 11, 2012; Accepted December 20, 2012)

**Abstract**

Semiparametric and nonparametric small area estimations have been studied to overcome a large variance due to a small sample size allocated in a small area. In this study, we investigate semiparametric and nonparametric mixed effect small area estimators using penalized spline and kernel smoothing methods respectively and compare their performances using labor statistics.

Keywords: Mixed effects model, nonparametric mixed effects model, semiparametric mixed effects model, kernel smoothing, penalized spline.

## 1. 서론

최근 관심이 집중되는 통계 분야의 하나로 소지역추정(small area estimation)이 부상하고 있다. 소지역 추정법은 지역 또는 도메인에 배분된 표본의 수가 작아 정확한 추정이 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.

흔히 소지역 추정법에 사용되고 있는 모형기반 추정법에서는 보조 정보의 양이 클수록 정확한 소지역 추정 방법이 가능하므로 보조 정보의 양을 증가시키는 방법이 연구되었다. 공간 정보를 이용하여 소지역 추정법의 정확도를 향상시키는 방법이 Kim 등 (2008)에 의해 연구되었으며 시계열 분석 모형을 이용한 방법 또한 연구되고 있다.

많은 학자들은 모수적 모형은 이미 완성 단계 혹은 한계에 이르렀다고 생각하고 있어 모수적 모형을 넘어선 다른 방법을 통해 모형의 정확도를 향상시키려는 움직임을 보이고 있으며 준모수 또는 비모수 소지역추정을 이용한 방법이 유망할 것으로 전망된다. 최근 들어 준모수적 방법이 소지역 추정법에 적용되기 시작하였는데, Opsomer 등 (2008), Salvati 등 (2010)이 준모수적 방법을 이용한 소지역 추정에 관하여 연구 결과를 발표하였다. 이들은 선형혼합모형을 스플라인 평활을 이용해 비모수적 방법으로 확장한 것이다. 최근 Jeong과 Shin (2012)는 커널 평활을 이용해 기존의 선형혼합모형에 기초한 모형기반 소지역 추정법을 비모수적 모형으로 확장하는 방법을 제시하고 그 유효성을 실증하였다. 그러나 이 방

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

법은 계산량이 많고 평활량을 결정해야 하는 등 실무에 적용하기에 어려운 점이 있다. 이에 본 논문에서는 Opsomer 등 (2008)이 제안한 방법인 준모수 소지역 추정량을 살펴보고, 이 방법과 Jeong과 Shin (2012)의 비모수 소지역 추정법의 결과를 비교하였다.

본 논문은 다음과 같이 구성되었다. 2절에서 선형혼합모형을 이용한 모형기반 소지역 추정법을 간단히 설명하고, Jeong과 Shin (2012)가 제안한 비모수 혼합 소지역추정법을 간단히 살펴보았다. 또한 벌점 스플라인을 이용한 준모수 혼합 소지역추정법을 살펴보았다. 3절에서는 실제 자료 분석을 통하여 각 추정량의 우수성을 비교하였으며, 4절에서 결론을 맺는다.

## 2. 비모수혼합모형을 이용한 소지역 추정

본 논문에서는 소지역 추정에서 사용되는 지역수준자료(area level data)와 단위수준자료(unit level data) 중에서 단위수준자료에 관하여 연구하였다. 또한 전 논문을 통하여 다음과 같은 설계를 사용하였다. 크기가 $N$인 모집단 $U$가 $d$개의 소지역의 모집단 $U_j$, $j = 1, 2, \ldots, d$으로 구성되어 있다고 하자. $j$번째 소지역의 모집단 $U_j$의 크기를 $N_j$라 하면 $\sum_{j=1}^{d} N_j = N$가 된다. 연속형인 관심변수 $y$에 대해 크기 $n$인 표본을 추출하되 각 소지역에서 얻어진 표본 크기를 $n_j$라 하면, $\sum_{j=1}^{d} n_j = n$가 된다. 또한 $s_j$를 $j$번째 소지역에서 추출된 표본 집합, $r_j$를 이 소지역에서 표본조사에서 제외된 집합이라 하면, $U_j = s_j \cup r_j$, $j = 1, 2, \ldots, d$이다. 본 논문의 연구 목적은 표본집합 $s_j$의 관심변수와 보조정보만을 이용하여 각 소지역 $U_j$에서 관심변수 $y$의 평균을 추정하는 것이다.

### 2.1. 선형혼합모형(linear mixed effect model)

$j$번째 소지역에서 $i$번째 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 $j = 1, 2, \ldots, d$는 지역을 $i = 1, 2, \ldots, n_j$는 $j$번째 지역내 관측값을 나타내는 인덱스이고, $\mathbf{x}_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijp})^T$는 보조변수 벡터, $\mathbf{z}_{ij} = (z_{ij1}, z_{ij2}, \ldots, z_{ijp})^T$는 지역관련 보조변수 벡터, $\boldsymbol{\beta}$는 고정효과(fixed effects)로서 보조정보 $x_{ij}$에 관련된 모수 벡터, $\boldsymbol{\gamma}_j \sim N(\mathbf{0}_q, \mathbf{G})$는 지역에 따른 랜덤효과(area-specific random effect), $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$는 오차항이다. $\mathbf{0}_m$은 모든 성분이 0이고 길이가 $m$인 벡터를, $\mathbf{1}_m$은 모든 성분이 1이고 길이가 $m$인 벡터를 나타낸다. 이를 각 소지역별로 묶어 행렬 및 벡터 기호로 나타내면 다음과 같다.

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

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

이들을 다시 각 소지역에 대해 열방향으로 쌓아올려 단번에 나타내면

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon} \tag{2.2}$$

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

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

이 된다. 여기서 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.

이때 랜덤효과 $\boldsymbol{\gamma}$와 오차항 $\boldsymbol{\epsilon}$이 서로 독립임을 가정하면 최량선형불편예측 이론에 의해

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

가 $\beta$와 $\gamma_j$의 최량선형불편예측(BLUP)의 추정값이 된다 (McCullough와 Searle 2001). 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 분산공분산 행렬 $R$과 $G$는 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 소지역 추정량은 다음과 같다.

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

이상에 관한 자세한 내용은 Rao (2003)을 참조하기 바란다.

### 2.2. 비모수혼합모형(nonparametric mixed effects model)

이 절에서는 다음과 같은 비모수혼합모형에 기반한 소지역 추정법을 설명하였다.

$$y_{ij} = \eta(x_{ij}) + \gamma(x_{ij}) + \epsilon_{ij}, \quad i = 1, 2, \ldots, n_j, \ j = 1, 2, \ldots, d, \tag{2.3}$$

여기서 $\eta(\cdot)$는 고정효과를 나타내는 함수, $\gamma_j(\cdot)$는 소지역 $j$에 해당하는 랜덤효과를 나타내는 함수이다. 문제를 단순화하기 위하여 보조변수들은 단변량인 경우로 한정해 설명하였다. 함수 $\eta$와 $\gamma_j$가 충분히 매끈하면 $x_{ij}$가 $x$ 근방에 있을 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. 따라서 식 (2.3)의 비모수적혼합모형은 다음과 같은 선형혼합모형으로 근사가 가능하다. 즉 $x_{ij} \approx x$일 때

$$\begin{aligned} y_{ij} &= \widetilde{\mathbf{x}}_{ij}^T\widetilde{\boldsymbol{\beta}} + \widetilde{\mathbf{z}}_{ij}^T\widetilde{\boldsymbol{\gamma}} + \epsilon_{ij}, \\ \mathbf{y}_j &= \widetilde{\mathbf{X}}_j^T\widetilde{\boldsymbol{\beta}} + \widetilde{\mathbf{Z}}_j^T\widetilde{\boldsymbol{\gamma}} + \boldsymbol{\epsilon}_j, \\ \mathbf{y} &= \widetilde{\mathbf{X}}^T\widetilde{\boldsymbol{\beta}} + \widetilde{\mathbf{Z}}^T\widetilde{\boldsymbol{\gamma}} + \boldsymbol{\epsilon}, \end{aligned} \tag{2.4}$$

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

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

이 된다. 따라서 식 (2.3)의 비모수적혼합모형을 식 (2.1)에서 (2.2)의 선형혼합모형과 똑같은 형태로 근사시킬 수 있다 (Jeong과 Shin, 2012). 이제 $\widetilde{\boldsymbol{\gamma}}_j \sim N(\mathbf{0}_{q+1}, \widetilde{\mathbf{G}})$을 가정하면 식 (2.4)의 로그우도함수가 얻어지며 이를 이용하면 지역별 평균 $\overline{Y}_j$, $j = 1, 2, \ldots, d$의 비모수적 소지역 추정량은

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

와 같이 얻어진다. 본 연구에서는 Jeong과 Shin (2012)과 같이 고정효과에 대해서는 국소선형($p = 1$), 랜덤효과에 대해서는 국소상수($q = 0$) 모형을 사용하였다. 또한 커널 $K$에 표준정규분포 $N(0, 1)$의 밀도함수인 $K(u) = (2\pi)^{-1/2}e^{-u^2/2}$을 사용했다. 여기서 국소로그우도를 최적화해 원하는 추정치 혹은 예측치는 SAS의 PROC MIXED나 R의 lme()와 같은 기존의 통계 계산 프로그램을 적용하면 쉽게 얻을 수 있다 (Wu와 Zhang, 2006; Jeong과 Shin, 2012).

### 2.3. 스플라인평활법을 이용한 준모수적 소지역추정법

식 (2.1)의 $f(x_{ij})$에 대해 아래와 같은 절사선형스플라인(truncated linear spline) 모형을 가정하자. 다만 수식을 간소화하기 위해 우선 $p = 1$인 경우에 한해 수식을 전개하였으며, 일반적인 $p > 1$인 경우는 쉽게 벡터로 일반화하여 표시할 수 있다.

$$f(x_{ij}) = \beta_0 + \beta_1 x_{ij} + \sum_{k=1}^{K} u_k(x_{ij} - \kappa_k)_+,$$

여기서 $\kappa_k$는 스플라인의 매듭점(knot)을 나타내는데 보통 관측된 $x_{ij}$ 값들의 분포의 분위수들로 정하며 매듭점의 개수 $K$는 표본 규모를 고려해 결정한다. 그러나 이 모형은 고정효과 부분을 과다적합(overfitting)하는 경향이 있으므로 모형의 복잡성에 대해 벌점을 주는 벌점회귀(penalized regression)을 이용해 스플라인 모수들을 추정한다. 즉 주어진 상수 $\lambda > 0$에 대해

$$\sum_{j=1}^{d}\sum_{i=1}^{n_j}\left\{y_{ij} - \beta_0 - \beta_1 x_{ij} - \sum_{k=1}^{K} u_k (x_{ij} - \kappa_k)_+\right\}^2 + \lambda \sum_{k=1}^{K} u_k^2$$

을 최소로 하는 $\beta$ 값들과 $u$ 값들을 추정하는 방식이다. 이상의 절차에 의한 추정 결과는 Ruppert 등 (2003)의 108쪽에서 제시한 바와 같이 $u$ 값들을 랜덤효과를 나타내는 계수인 것처럼 생각하고 BLUP 이론을 적용한 것과 같아지게 된다. 또한 위 과정 중에 벌점 모수 $\lambda$까지 함께 자료값에 의해 자동으로 결정되기 때문에 벌점 모수 결정 문제에 대해 고민할 필요가 없다는 장점이 있다. 이상의 내용을 식 (2.1)에 대입해 결합하면

$$y_{ij} = \bar{\boldsymbol{x}}_{ij}^T \bar{\boldsymbol{\beta}} + \mathbf{d}_{ij}^T \boldsymbol{u} + \mathbf{z}_{ij}^T \boldsymbol{\gamma}_j + \epsilon_{ij}$$

와 같은 선형혼합모형 형태로 다시 표현된다. 단, $\bar{\boldsymbol{x}}_{ij} = [1 \ x_{ij}]^T$, $\bar{\boldsymbol{\beta}} = [\beta_0 \ \beta_1]^T$, $d_{ij} = [(x_{ij} - \kappa_1)_+, (x_{ij} - \kappa_2)_+, \ldots, (x_{ij} - \kappa_K)_+]^T$, $\boldsymbol{u} = [u_1, u_2, \ldots, u_K]^T$이다. 따라서 앞 절에서 설명한 BLUP 이론을 랜덤효과 성분이 두 개인 경우로 확장 적용하면 고정효과 및 랜덤효과 성분의 모수들의 BLUP을 얻을 수 있고 이들을 이용해 원하는 소지역추정량을 얻을 수 있다. 이상의 내용은 Opsomer 등 (2008)의 267쪽 이하의 내용을 참조하기 바란다.

## 3. 2006 매월노동통계 자료를 이용한 모의실험

이 절에서는 본 논문에서 설명한 벌점스플라인을 이용한 준모수적 소지역 추정법과 Jeong과 Shin (2012)이 제안한 비모수적혼합모형 소지역추정량 그리고 기존의 선형혼합모형 소지역추정량의 성능을 비교하였다. 분석에 사용한 자료는 Jeong과 Shin (2012)에서와 같이 노동부의 '2006년 매월노동통계'의 원자료이다. 이는 Jeong과 Shin (2012)의 결과와 직접 비교하기 위한 것이다. 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. 각 소지역 내 표본 사업체 수를 $n_j$라 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) $U$를 생성한다. 생성된 의사모집단을 각 소지역별로 구별하여 $U_j$라 하고 그 크기 $N_j$를 구한다.

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

(3) 의사모집단 $U$로부터 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 $n_j$가 되도록 한다. 각 층에서 추출된 자료를 모은 것을 $s_j$라 하면 $r_j = U_j - s_j$이 된다.

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

**Table 3.1.** Simulation results

| 추정량 | $h$ | MSE($\times 10^9$) | MAE($\times 10^4$) | RE($\times 10^{-2}$) | ARE($\times 10^{-1}$) |
|---|---|---|---|---|---|
| 선형혼합 | $\infty$ | 1.3022 | 2.1199 | 1.9828 | 1.0163 |
| $\hat{\overline{Y}}_j^{MX}$ | (s.e) | (0.0004) | (0.0004) | (0.0014) | (0.0000) |
| 비모수혼합 | 0.10 | 0.1189 | 0.8051 | 0.5251 | 0.5200 |
| $\hat{\overline{Y}}_j^{NPMX}$ | (s.e) | (0.0006) | (0.0024) | (0.0030) | (0.0017) |
| 준모수혼합 | | 0.1526 | 0.8187 | 0.4222 | 0.4641 |
| $\hat{\overline{Y}}_j^{SPMX}$ | (s.e) | (0.0001) | (0.0002) | (0.0025) | (0.0014) |

(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 사용했다. 각 비교통계량의 구체적 형태는 다음과 같다.

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

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

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

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

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

다음의 Table 3.1은 500회의 반복실험을 통해 얻은 각종 비교통계량의 값들을 정리한 것이다. 비모수혼합모형의 경우 Jeong과 Shin (2012)의 결과 중에서 평활량 $h = 0.1$인 경우로 가장 우수한 결과를 수록하였다. 선형혼합모형 결과도 Jeong과 Shin (2012)의 결과와 동일하다. 이제 준모수혼합모형 결과와 선형혼합모형 결과와 비교하면 모든 비교 통계량에서 준모수혼합모형이 매우 우수한 결과를 주고 있음을 확인할 수 있다. 반면 비모수혼합모형과 준모수혼합모형 결롸는 비교 통계량에 따라 다른 결과를 주고있다. 즉 MSE의 경우 비모수혼합모형 결과가 우수한 반면 MAE, RE 그리고 ARE에서 준모수혼합모형의 결과가 우수한 것으로 나타났다. 따라서 준모수 방법과 비모수 방법의 우수성을 평가하기는 어렵다. 물론 비모수와 준모수 혼합모형이 선형혼합모형에 비해 매우 우수한 결과를 주는 것을 확인할 수 있다.

Figure 3.1은 두 모형의 적합정도를 확인하기 위하여 각각 선형혼합모형과 준모수혼합모형을 자료에 적합시킨 그림이다. 선형혼합모형의 경우 독립변수로 사용된 종사사 수에 직선으로 적합이 되었으나 준모수혼합모형의 경우 종사자 수가 1,000명에서 3,000명인 구간에서는 직선이 아닌 곡선으로 적합되어 차이를 보이고 있음을 확인할 수 있으며 이후 직선의 형태를 유지하나 두 모형의 기울기에 차이가 있음을 확인할 수 있다. 또한 각 모형의 잔차를 확인하기 위하여 각 모형에서 구한 잔차를 Figure 3.2에 나타내

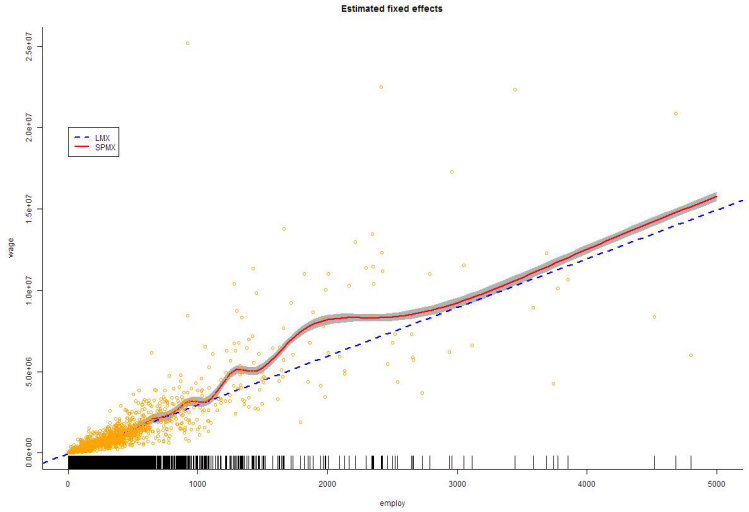


**Figure 3.1.** Comparison of estimated fixed effect compon

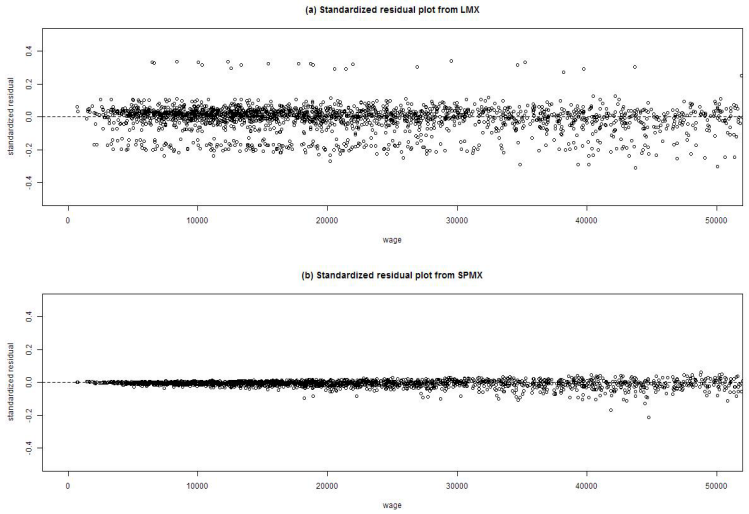


**Figure 3.2.** Comparison of standardized residuals

었다. 위의 그림이 선형혼합모형의 표준화잔차 그림이고 아래 그림이 준모수혼합모형의 표준화잔차 그림으로 두 그림을 비교하면 준모수혼합모형의 표준화잔차가 작은 것을 확인할 수 있다. 따라서 그림을 통한 비교에서도 준모수혼합모형이 본 자료에 더욱 적합한 것을 확인 할 수 있다.

## 4. 결론 및 전망

소지역 추정 방법으로 모형기반 추정법에 대한 관심이 증대되는 가운데 준모수 벌점회귀 또는 비모수 함수추정 기법은 이 분야에서 활용도가 매우 높을 것으로 기대된다. Jeong과 Shin (2012)이 제안한 국소

다항모형을 이용한 비모수적혼합모형 소지역 추정법을 2006 매월노동통계 자료에 대해 적용한 결과 매우 우수한 성능을 보였다. 비모수적 방법 고유의 유연성을 고려할 때 여타 자료에서도 여전히 우수한 성능을 보일 것으로 기대된다. 그러나 일반적으로 비모수적함수추정 기법을 적용할 때 늘 그렇듯이 비모수적 혼합모형 적합 시 평활 모수(smoothing parameter)를 적절히 선택해야 하는 어려움이 있다. 이에 반하여 준모수 벌점회귀를 이용한 소지역추정법은 최적의 평활량을 선택하는 어려움이 없으며 이 모형에서 사용하는 벌점 모수 $\lambda$는 자료에서 자동으로 추정할 수 있어 사용에 매우 편리한 방법이다. 또한 모의실험 결과에서도 알 수 있듯이 비모수 소지역추정법에 비해 그 우수성이 떨어지지 않는다. 따라서 현장에서 이 방법을 사용하게 되면 우수한 소지역 결과를 얻을 수 있을 것으로 기대된다.

## References


Jeong, S.-O. and Shin, K.-I. (2012). Small area estimation via nonparametric mixed effects model, *The Korean Journal of Applied Statistics*, **25**, 457–464.

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.

McCulloch, C. and Searle, S. (2001). *Generalized, Linear and Mixed Models*, Wiley.

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.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*, Cambridge.

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.

# 비모수와 준모수 혼합모형을 이용한 소지역 추정

정석오[a] · 신기일[a,1]

[a]한국외국어대학교 통계학과

(2012년 10월 12일 접수, 2012년 12월 11일 수정, 2012년 12월 20일 채택)

**요 약**

지역 또는 도메인에 작은 크기의 표본이 배정되어 추정의 정도가 나쁜 경우에 사용되는 준모수적 또는 비모수적 소지역 추정법은 최근 많은 연구가 진행되고 있다. 본 논문에서는 커널을 이용한 국소다항 혼합모형 소지역 추정법과 벌점 스플라인을 이용한 혼합모형 소지역 추정법이 연구되었다. 이 두 방법과 소지역추정에 흔히 사용되고 있는 선형 혼합모형을 모의실험을 통해 그 우수성을 비교하였다.

주요용어: 선형혼합모형, 비모수적 혼합모형, 커널 평활법, 준모수적 혼합모형, 벌점스플라인.

이 논문은 2010년도 정부(교육과학기술원)의 재원으로 한국연구재단의 기초연구사업 (2010-0008807, 신기일)과 한국외국어대학교 교내연구비의 지원을 받아 수행되었음 (2012, 정석오).

[1]교신저자: (449-791) 한국외국어대학교 통계학과, 경기도 용인시 모현면 산 89, 교수.
E-mail: keyshin@hufs.ac.kr