The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.4.605
27(4), 605–617

Shrinkage Small Area Estimation Using a Semiparametric Mixed Model

Seok-Oh Jeong[a] · Manho Choo[a] · Key-Il Shin[a,1]

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

(Received March 20, 2014; Revised June 10, 2014; Accepted June 18, 2014)

Abstract

Small area estimation is a statistical inference method to overcome large variance due to a small sample size allocated in a small area. A shrinkage estimator obtained by minimizing relative error(RE) instead of MSE has been suggested. The estimator takes advantage of good interpretation when the data range is large. A semiparametric estimator is also studied for small area estimation. In this study, we suggest a semiparametric shrinkage small area estimator and compare small area estimators using labor statistics.

Keywords: Spline regression, linear mixed estimation, empirical best linear unbiased predictor, shrinkage estimator.

1. 서론

소지역추정을 간단히 정리하면 지역 또는 도메인에 배분된 표본의 수가 작아 정확한 추정이 불가능할 때 이를 극복하는 통계적 방법이다. 소지역추정법은 크게 자료기반 또는 설계기반(data based, design-based) 추정법과 모형기반(model-based) 추정법으로 나누어진다. 설계기반 추정법은 얻어진 자료만을 사용하므로 추가적인 정보를 사용하지 않기 때문에 추정의 정확도를 향상시키기에는 한계가 있다. 따라서 최근 분석에서는 추가적인 정보, 즉 보조정보(auxiliary information)를 이용하여 좀 더 정확한 추정이 가능한 모형기반 추정법이 주로 사용되고 있다. 모형기반 추정법으로는 기본적으로 회귀추정법, 비추정법과 같은 일반회귀추정법(generalized regression method)이 사용된다. 또한 선형혼합모형, 경험적베이즈추정법, 계층적베이즈추정법 등도 사용된다. 이러한 모든 모형들은 모수적 모형을 따르는 것으로 이미 많은 소지역 추정에 사용되고 있다. 자세한 내용은 Rao (2003)을 참조하기 바란다.

특히 보조 정보의 양이 클수록 정확한 소지역 추정 방법이 가능하므로 보조 정보의 양을 증가시키는 방법이 최근 연구되었다. 예를 들어 공간정보를 이용하여 소지역추정법을 향상시키는 방법이 연구되었으며 시계열 분석 모형을 이용한 방법 또한 연구되고 있다. 다른 한편으로는 추정의 정확도를 나타내는 기준에 관한 연구가 진행되었다. 일반적으로 사용되는 기준은 MSE(mean squared error)이다. 이 기준은 여러 통계 분야에서 기본적으로 사용되고 있다. 그러나 우리나라의 소지역처럼 지역의 크기 또는 도메인의 크기에 차이가 있는 경우에는 상대오차(RE; relative error)를 사용하는 것이 타당한 경우도 있

The research was supported by Hankuk University of Foreign Studies research fund(2014).
[1]Corresponding author: Department of Statistics, Hankuk University of Foreign Studies, Yongin, Gyeonggi 449-791, Korea. E-mail: keyshin@hufs.ac.kr

다. 이 기준을 적용하게 되면 소지역 추정량은 축소소지역추정량(shrinkage small area estimator)이 되며 Hwang과 Shin (2008)은 이에 관하여 연구하였다.

다른 한편으로는 모수적 모형개발 이외의 방법인 비모수적 함수 추정을 이용한 소지역 추정법이 연구되었다. Opsomer 등 (2008)은 준모수적 방법을 이용한 소지역 추정에 관하여 연구하였다. 이 방법은 모수적 방법에서 가장 흔하게 사용되는 방법인 선형혼합모형에 준모수적인 방법을 접목한 방법이다. 또한 Salvati 등 (2010)도 비모수회귀를 이용한 소지역 추정에 관한 연구를 수행하였다. 또한 Jeong과 Shin (2013)은 비모수와 준모수혼합모형을 이용한 소지역 추정을 연구하였다.

이미 여러 비모수적 방법을 이용한 소지역 추정법이 연구되었으나 비모수적 방법을 이용한 소지역 추정법 중에서 상대오차를 최소로 하는 비모수적 소지역 추정법은 상대적으로 연구가 미미한 상태이다. 최근 상대오차를 비모수회귀에 적용한 논문인 Jones 등 (2008)이 발표되었으며, Jeong과 Shin (2008)은 Jones 등 (2008)의 평활량 선택과 경계점 문제를 보완한 논문을 발표하였다. 그러나 Jeong과 Shin (2013)의 연구 결과를 살펴보면 비모수적추정에 비해 준모수적추정이 사용이 간단하면서도 효율성이 우수한 것을 확인 할 수 있다. 이에 본 연구에서는 준모수혼합모형에 축소추정법을 결합한 소지역추정법을 연구하였다.

본 논문은 2절에서 현재 사용되고 있는 소지역추정량 중 가장 많이 사용되는 선형혼합추정량과 준모수혼합모형을 이용한 소지역 추정량과 축소추정량에 관해 간단히 설명하였다. 또한 기존의 혼합모형과 축소소지역추정량을 결합한 새로운 준모수축소소지역추정량을 제안하였다. 3절에서는 Lee 등 (1995)의 모집단 생성 방법으로 만들어진 자료를 이용해 모의실험이 수행되었다. 4절에서는 매월노동통계 자료를 이용한 사례연구를 통하여 기존의 추정량과 제안된 추정량의 우수성을 비교했으며 5절에 결론이 있다.

2. 혼합모형을 이용한 축소소지역추정법

모수적 소지역 추정법은 설계기반 추정법과 모형기반 추정법으로 나누어진다. 서론에서 언급한 것처럼 최근의 분석에서는 모형기반 추정법이 주로 사용되므로 모형기반 추정법을 살펴보았다. 특히 모형기반 추정법 중에서 연속형 변수에서 가장 많이 사용하고 있는 선형혼합모형 방법을 2.1절에서 살펴보았다. 또한 준모수 혼합모형을 이용한 소지역 추정법을 2.2절에서 설명하였다. 이 내용은 Jeong과 Shin (2013)에도 자세히 설명되어 있어 본 논문에서는 간단히 설명하였다. 또한 2.3절에서는 본 연구에서 제안한 상대오차를 이용한 축소추정법을 설명하였다.

본 연구에서는 소지역 추정에서 사용되는 지역수준자료(area level data)와 단위수준자료(unit level data) 중에서 단위수준자료가 사용되었다. 또한 전 연구를 통하여 다음과 같은 설계를 사용하였다.

크기가 N인 모집단 U가 d개의 소지역 모집단 U_j, $j = 1, 2, \ldots, d$로 구성되어 있다고 하자. j번째 소지역 모집단 U_j의 크기를 N_j라 하면 $\sum_{j=1}^{d} N_j = N$이 된다. 연속형인 관심변수 y에 대해 크기가 n인 표본을 추출하되 각 소지역에서 얻어진 표본크기를 n_j라 하면 $\sum_{j=1}^{d} n_j = n$이 된다. 또한 s_j를 j번째 소지역에서 추출된 표본집합, r_j를 이 소지역에서 표본조사에서 제외된 집합이라 하면 $U_j = s_j \cup r_j$이다. 본 연구의 목적은 표본집합 s_j의 관심변수와 보조정보만을 이용하여 각 소지역 U_j에서 관심변수 y의 평균을 추정하는 것이며 표본조사에서 제외된 집합인 r_j는 보조정보가 있다는 가정을 사용하였다.

2.1. 선형혼합모형(linear mixed effects model)

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

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

여기서 $\boldsymbol{x}_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijp})^T$와 $\boldsymbol{z}_{ij} = (z_{ij1}, z_{ij2}, \ldots, z_{ijq})^T$는 보조변수 벡터, $\boldsymbol{\beta}$는 고정효과(fixed effects), $\boldsymbol{\gamma_j} \sim N(\mathbf{0}_q, \boldsymbol{G})$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 m인 벡터를, $\mathbf{1}_m$은 모든 성분이 1이고 길이가 m인 벡터를 나타낸다.

이를 각 소지역별로 묶어 행렬 및 벡터 기호로 나타내면 다음과 같다.

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

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

이들을 다시 각 소지역에 대해 열 방향으로 쌓아올려 행렬로 나타내면

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

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

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

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

랜덤효과 $\boldsymbol{\gamma}$와 오차항 $\boldsymbol{\epsilon}$이 서로 독립임을 가정하고 각 분산성분 $\boldsymbol{G}$와 $\boldsymbol{R}$이 주어진 경우 로그우도함수는

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

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

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y},$$
$$\hat{\boldsymbol{\gamma}}_j = \boldsymbol{G}\boldsymbol{Z}_j^T\boldsymbol{V}_j^{-1}\left(\boldsymbol{y}_j - \boldsymbol{X}_j\hat{\boldsymbol{\beta}}\right), \quad j = 1, 2, \ldots, d$$

와 같다. 단, $\boldsymbol{V}_j = \boldsymbol{Z}_j\boldsymbol{G}\boldsymbol{Z}_j^T + \boldsymbol{R}_j$, $\boldsymbol{V} = \mathrm{diag}(\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_d) = \boldsymbol{Z}\bar{\boldsymbol{G}}\boldsymbol{Z}^T + \boldsymbol{R}$이다. 이 추정과정에서 필요한 분산공분산행렬 $\boldsymbol{R}$과 $\boldsymbol{G}$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 $\boldsymbol{x}_{ij}^T\hat{\boldsymbol{\beta}} + \boldsymbol{z}_{ij}^T\hat{\boldsymbol{\gamma}}_j$와 같이 구할 수 있고, 따라서 지역별 평균 $\bar{Y}_j$, $j = 1, 2, \ldots, d$의 소지역 추정량은 다음과 같이 얻어진다.

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

이상에 관한 자세한 내용은 Rao (2003)을 참조하기 바란다.

2.2. 스플라인평활법을 이용한 준모수혼합 소지역추정법

식 (2.1)의 $y_{ij} = f(\boldsymbol{x}_{ij})$에 대해 아래와 같은 절사선형스플라인(truncated linear spline) 모형을 가정하자. 이 절에서는 수식을 간소화하기 위해 $p = 1$인 경우에 한해 수식을 전개하였으며 $p > 1$인 경우는 쉽게 벡터로 일반화하여 표시될 수 있다.

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

여기서 κ_k는 스플라인의 매듭점(knot)을 나타내는데 보통 관측된 x_{ij}값들의 분포의 분위수들로 정하며 매듭점의 개수 K는 표본 규모를 고려해 결정한다. 여기서 $(\cdot)_+$는 괄호안의 수가 양수이면 그 값이 되고, 음수이면 '0'이 되는 것을 의미한다. 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

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

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

$$y_{ij} = \bar{\boldsymbol{x}}_{ij}^\top\bar{\boldsymbol{\beta}} + \boldsymbol{d}_{ij}^\top\boldsymbol{u} + \boldsymbol{z}_{ij}^\top\boldsymbol{\gamma}_j + \epsilon_{ij} \tag{2.6}$$

와 같은 선형혼합모형 형태로 다시 표현된다. 단, $\bar{\boldsymbol{x}}_{ij} = [1 \quad x_{ij}]^\top$, $\bar{\boldsymbol{\beta}} = [\beta_0 \quad \beta_1]^\top$, $\boldsymbol{d}_{ij} = [(x_{ij} - \kappa_1)_+, (x_{ij} - \kappa_2)_+, \ldots, (x_{ij} - \kappa_K)_+]^\top$, $\boldsymbol{u} = [u_1, u_2, \ldots, u_K]^\top$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. 따라서 준모수혼합모형을 이용한 소지역 추정량은 다음과 같다.

$$\hat{\bar{Y}}_j^{SPMX} = N_j^{-1}\left\{\sum_{i\in s_j} y_{ij} + \sum_{i\in r_j}\left(\bar{\boldsymbol{x}}_{ij}^T\hat{\bar{\boldsymbol{\beta}}} + \boldsymbol{d}_{ij}^T\hat{\boldsymbol{u}} + \boldsymbol{z}_{ij}^T\hat{\boldsymbol{\gamma}}_j\right)\right\}. \tag{2.7}$$

이상의 내용은 Opsomer 등 (2008)의 267쪽 이하의 내용과 Jeong과 Shin (2013)을 참조하기 바란다.

2.3. 상대오차를 이용한 소지역추정법

상대오차를 이용한 추정법은 Park과 Stefanski (1997)에서 연구되었고 이를 이용한 축소소지역추정법은 Hwang과 Shin (2008)에서 연구되었다. 이 절에서는 상대오차를 이용하여 비모수적으로 추정하는 방법을 간단히 살펴보았다. 먼저 일반적인 MSE를 최소로 하여 얻은 추정량은 다음과 같다.

$$\arg\min_{\tilde{Y}} E\left(\left(Y-\tilde{Y}\right)^2|X=x\right)=E\left(Y|X=x\right)=\mu(x).$$

그러나 상대오차를 고려하는 경우 다음과 같은 결과를 얻을 수 있다.

$$g(x)=\arg\min_{\tilde{Y}} E\left(\frac{\left(Y-\bar{Y}\right)}{\bar{Y}}\right)^2=\frac{E\left(Y^{-1}|X=x\right)}{E\left(Y^{-2}|X=x\right)}.$$

이제 Hwang과 Shin (2008)의 결과를 이용하면 다음 식을 얻는다.

$$\frac{E\left(Y^{-1}|X=x\right)}{E\left(Y^{-2}|X=x\right)}\approx\mu(x)\frac{1+\mathrm{CV}(x)^2}{1+3\mathrm{CV}(x)^2}. \tag{2.8}$$

이제 식 (2.8)의 결과를 본 연구의 결과인 식 (2.5)와 식 (2.7)에 적용하게 되면 다음의 축소소지역추정결과를 얻게 된다.

$$\bar{g}(x)=\hat{\mu}(x)\frac{1+\widehat{\mathrm{CV}}(x)^2}{1+3\widehat{\mathrm{CV}}(x)^2}. \tag{2.9}$$

따라서 $\hat{\mu}(x)$에 식 (2.5) 또는 식 (2.7)을 대입하게 되면 원하는 축소소지역추정량이 얻어진다.

2.4. 제안된 축소소지역추정량

축소소지역추정량은 식 (2.9)에 기초하여 얻어지며 식 (2.9)에 포함된 $\hat{\mu}(x)$는 주어진 모형에 따라 다르게 얻어진다. 먼저 선형혼합모형을 이용할 경우에는 식 (2.5)를 이용하여 얻어진다. 즉

$$\hat{\mu}(x)=\hat{\bar{Y}}_j^{MX}=N_j^{-1}\left\{\sum_{i\in s_j}y_{ij}+\sum_{i\in r_j}\left(\boldsymbol{x}_{ij}^T\hat{\boldsymbol{\beta}}+\boldsymbol{z}_{ij}^T\hat{\boldsymbol{\gamma}}_j\right)\right\}$$

이 되고, 얻어진 추정량을 이용하여 $\widehat{\mathrm{CV}}(x)$도 추정하게 된다. 같은 방법으로 식 (2.7)의

$$\hat{\mu}(x)=\hat{\bar{Y}}_j^{SPMX}=N_j^{-1}\left\{\sum_{i\in s_j}y_{ij}+\sum_{i\in r_j}\left(\bar{\boldsymbol{x}}_{ij}^T\hat{\boldsymbol{\beta}}+\boldsymbol{d}_{ij}^T\hat{\boldsymbol{u}}+\boldsymbol{z}_{ij}^T\hat{\boldsymbol{\gamma}}_j\right)\right\}$$

을 이용하게 되면 준모수혼합모형을 이용한 축소소지역추정량이 구해진다. 축소추정량을 계산하기 위해서는 $\widehat{\mathrm{CV}}(x)$의 계산이 필요하며 이때 분산 추정량 $\hat{\sigma}^2$은 붓스트랩 방법을 이용하여 계산할 수 있다. 즉 표본으로 추출된 자료, y_{ij}, $i\in s_j$를 복원추출로 표본 수만큼 랜덤하게 R회 반복하여 추출한 후 얻어진 $\hat{\mu}(x)^{(r)}, r=1,\ldots,R$을 이용하여 분산을 추정하면 된다. 즉

$$\hat{\sigma}^2=\frac{1}{R-1}\sum_{r=1}^{R}\left(\hat{\mu}(x)^{(r)}-\bar{\hat{\mu}}(x)\right)^2$$

으로 계산하며 $\widehat{\mathrm{CV}}(x)^2 = \hat{\sigma}^2/\hat{\mu}(x)^2$로 얻어진다. 따라서 $\hat{\bar{Y}}_j^{SH\text{-}MX}$, $\hat{\bar{Y}}_j^{SH\text{-}SPMX}$을 각각 축소선형혼합 소지역추정량과 축소준모수혼합 소지역추정량이라 하면 다음과 같이 얻어진다.

$$\hat{\bar{Y}}_j^{SH\text{-}MX} = \hat{\bar{Y}}_j^{MX} \times \frac{1 + \widehat{\mathrm{CV}}^{MX}(x)^2}{1 + 3\widehat{\mathrm{CV}}^{MX}(x)^2}, \tag{2.10}$$

$$\hat{\bar{Y}}_j^{SH\text{-}SPMX} = \hat{\bar{Y}}_j^{SPMX} \times \frac{1 + \widehat{\mathrm{CV}}^{SPMX}(x)^2}{1 + 3\widehat{\mathrm{CV}}^{SPMX}(x)^2}. \tag{2.11}$$

3. 모의실험

3.1. 모집단 자료 생성

모의실험을 위한 자료의 생성과정은 Lee 등 (1995)에서 사용한 동일한 방법을 사용하였다. 먼저 크기 N = 50,000인 모집단을 다음과 같이 생성하였다. 보조자료 x_k는 평균 48이고 분산 768을 갖는 감마분포로부터 생성하였다. 주어진 x_k값에 대하여 모두 네 종류의 조사자료 y_k를 발생시켰는데, 각각 평균 $\mu(x) = a + bx + cx^2$이고 분산 $\sigma^2(x) = d^2x^{2g}$을 갖는 감마분포를 가정하였다. 이는 보조자료 x_k가 커질수록 분산이 커지도록 변수를 생성하였는데 이러한 현상은 현실적인 사업체 자료에서는 매우 흔한 현상이므로 이를 반영하기 위해서이다. Table 3.1은 선택된 상수 a, b, c, d, g의 값을 나타낸다. 첫 번째로 생성된 자료는 보조변수와의 관계가 원점을 지나는 비례적 형태(ratio)이고, 두 번째 자료는 양의 절편 값을 갖는 선형관계(regression)를 갖도록 하였다. 또한 볼록형과 오목형의 관계가 성립하도록 상수를 조정한 후 자료를 생성하였다.

다음으로 생성된 50,000개 자료를 50개의 소지역으로 나누었다. 이를 위해 먼저 관심변수 y_k와 보조변수 x_k를 오름차순으로 정렬한 후, 크기순으로 각 소지역 층에 자료를 배정하였다. 층의 크기는 22개 층이 250개에서 700개의 원소를 갖고 있으며, 6개 층이 800개에서 1,000개 그리고 나머지 22개 층이 1,100개에서 1,750개의 원소를 갖고 있다. 이는 매우 흔하게 소지역의 크기가 일정하지 않기 때문에 이를 반영하기 위해서이다.

생성된 자료에 식 (2.5), 식 (2.7), 식 (2.10) 그리고 식 (2.11)을 적용하여 소지역추정량을 얻었다. 각 추정량의 우수성을 살펴보기 위하여 500번의 반복이 이루어졌다. 또한 분산 추정을 위한 붓스트랩 반복수 R = 200을 사용하였다. 또한 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 고려하였다. 각 비교통계량의 구체적 형태는 다음과 같다.

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

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

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

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

Table 3.1. Coefficients used for generating data

y_k 형태	a	b	c	d	g
비례형(ratio)	0	1.50	0.00	5.13	0.50
선형(regression)	20	1.50	0.00	13.79	0.25
볼록형(convex)	0	0.25	0.01	4.91	0.50
오목형(concave)	0	3.00	−0.01	5.60	0.50

Table 3.2. The result from ratio data($R = 500, B = 200$)

추정량	MSE	MAE	RE($\times 10^{-3}$)	ARE($\times 10^{-2}$)
선형혼합 $\hat{\bar{Y}}_j^{MX}$	1.3969	0.9064	1.6494	2.5380
	(0.0258)	(0.0078)	(0.0564)	(0.0318)
준모수혼합 $\hat{\bar{Y}}_j^{SPMX}$	1.4271	0.9043	1.8328	2.5670
	(0.0255)	(0.0073)	(0.0680)	(0.0323)
축소선형혼합 $\hat{\bar{Y}}_j^{SH\text{-}MX}$	1.4173	0.9130	1.6957	2.5718
	(0.0265)	(0.0081)	(0.0633)	(0.0350)
축소준모수혼합 $\hat{\bar{Y}}_j^{SH\text{-}SPMX}$	1.4479	0.9109	2.1133	2.6330
	(0.0261)	(0.0075)	(0.1016)	(0.0369)

Table 3.3. The result from regression data($R = 500, B = 200$)

추정량	MSE	MAE	RE($\times 10^{-3}$)	ARE($\times 10^{-2}$)
선형혼합 $\hat{\bar{Y}}_j^{MX}$	1.6069	1.0186	0.6131	1.7732
	(0.0242)	(0.0073)	(0.0112)	(0.0136)
준모수혼합 $\hat{\bar{Y}}_j^{SPMX}$	1.5706	1.0059	0.6010	1.7522
	(0.0228)	(0.0068)	(0.0110)	(0.0129)
축소선형혼합 $\hat{\bar{Y}}_j^{SH\text{-}MX}$	1.6431	1.0301	0.6403	1.8040
	(0.0254)	(0.0076)	(0.0125)	(0.0147)
축소준모수혼합 $\hat{\bar{Y}}_j^{SH\text{-}SPMX}$	1.5862	1.0108	0.6149	1.7661
	(0.0230)	(0.0069)	(0.0115)	(0.0131)

여기서 $\bar{Y}_j$는 j번째 소지역의 참값이고, $\hat{\bar{Y}}_j^{(r)}$은 r번째($r = 1, 2, \ldots, R$) 모의실험에서 얻은 소지역 추정값을 뜻한다.

3.2. 모의실험 결과

모의실험 결과는 Table 3.2 ∼ Table 3.5에 수록하였다. 먼저 Table 3.2의 비례형 결과를 살펴보면 선형혼합모형과 준모수혼합모형 결과는 큰 차이를 보이고 있지 않으나 선형혼합모형의 결과가 준모수혼합모형의 결과에 비해 우수한 것을 확인할 수 있다. 특히 RE를 기준으로 할 때 선형혼합모형의 결과가 매우 우수하다. 또한 축소선형모형을 적용한 결과는 모두 좋지 않는 것으로 확인되었다. 다음의 선형자료 결과인 Table 3.3을 살펴보면 역시 선형혼합모형과 준모수혼합모형이 큰 차이를 보이고 있지 않지만 선형혼합 모형에 비해 준모수혼합모형의 결과가 모든 통계량에서 약간 우수한 결과를 주고 있다. 이는 비례형과 선형 모두 직선관계를 유지하고 있어 두 방법 모두 적합하기 때문이다. 그러나 축소선형혼합모형 또는 축소준모수혼합모형을 사용한 경우에는 결과가 좋지 않게 나온다.

이제 독립변수와 종속변수가 직선관계가 아닌 경우인 볼록형과 오목형 결과를 살펴보자. 먼저 볼록형 자료의 결과인 Table 3.4를 살펴보면 선형혼합모형에 비해 준모수혼합모형의 결과가 매우 우수한 것을 확인할 수 있다. 이는 물론 두 변수간의 관계가 선형이 아니기 때문에 발생할 수 있는 자연스러운 결과

Table 3.4. The result from convex data($R = 500, B = 200$)

추정량	MSE	MAE	RE($\times 10^{-1}$)	ARE($\times 10^{-1}$)
선형혼합 $\hat{\bar{Y}}_j^{MX}$	7.7467	2.2229	10.3679	3.4252
	(0.0763)	(0.0111)	(0.4741)	(0.0466)
준모수혼합 $\hat{\bar{Y}}_j^{SPMX}$	2.0774	1.0891	7.7244	2.1662
	(0.0317)	(0.0071)	(0.3378)	(0.0378)
축소선형혼합 $\hat{\bar{Y}}_j^{SH\text{-}MX}$	7.8188	2.2823	7.4634	3.2473
	(0.0734)	(0.0113)	(0.4029)	(0.0446)
축소준모수혼합 $\hat{\bar{Y}}_j^{SH\text{-}SPMX}$	2.0383	1.0878	4.5412	1.9054
	(0.0293)	(0.0067)	(0.2314)	(0.0310)

Table 3.5. The result from concave data($R = 500, B = 200$)

추정량	MSE($\times 10^3$)	MAE($\times 10^1$)	RE($\times 10^{-4}$)	ARE($\times 10^{-2}$)
선형혼합 $\hat{\bar{Y}}_j^{MX}$	5.6669	5.7986	4.9291	1.1531
	(0.0995)	(0.0490)	(0.2261)	(0.0145)
준모수혼합 $\hat{\bar{Y}}_j^{SPMX}$	1.6880	3.1018	1.9977	0.6556
	(0.0481)	(0.0348)	(0.1451)	(0.0133)
축소선형혼합 $\hat{\bar{Y}}_j^{SH\text{-}MX}$	5.6623	5.8043	4.9246	1.1558
	(0.0975)	(0.0489)	(0.2176)	(0.0142)
축소준모수혼합 $\hat{\bar{Y}}_j^{SH\text{-}SPMX}$	1.6629	3.0877	1.8863	0.6463
	(0.0466)	(0.0345)	(0.1378)	(0.0129)

이기도 하다. 따라서 직선 관계가 아닌 경우에는 당연하지만 준모수혼합모형을 사용할 필요가 있다. 다음으로 준모수혼합모형과 축소준모수혼합모형을 비교해 보면 축소준모수혼합모형의 결과가 모든 통계량을 기준으로 할 때 좋게 나온 것을 확인할 수 있다. 특히 축소모형을 사용하는 주된 목적이 상대오차 즉 RE를 최소로 하는 것인데 이 경우에는 매우 크게 상대오차를 줄일 수 있다. 따라서 볼록형 자료인 경우에는 축소모형을 사용할 필요가 있다.

또한 오목형 자료 결과인 Table 3.5 결과를 살펴보면 역시 준모수혼합모형을 사용한 결과가 우수한 결과를 주고 있다. 또한 축소준모수혼합모형을 사용한 결과를 살펴보면 모든 비교 통계량에서 우수한 결과를 주고 있음을 확인할 수 있다. 특히 축소준모수혼합모형의 상대오차는 크게 감소하였다.

4. 2006 매월노동통계 자료를 이용한 사례분석

이 절에서는 본 연구에서 제안한 준모수혼합축소 소지역추정량의 성능을 기존의 선형혼합모형 소지역추정량과 비교하였다. 분석에 사용한 자료는 노동부의 '2006년 매월노동통계'의 원자료이다. 이 자료는 전국의 $n = 7{,}038$ 사업체를 조사해 얻은 임금총액(y) 및 종사자 수(x)의 자료이며, 소지역은 전국의 $d = 47$개 지청이 된다. 각 소지역 내 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 실시하였다. 다음의 절차는 최근 논문인 Salvati 등 (2010), Jeong과 Shin (2013)의 모의실험 방법을 사용한 것이다.

(1) 크기가 7,038인 원자료에 대해 10회의 재추출(복원 허용)을 실시해 크기가 ($N = 7{,}038 \times 10$)인 의사모집단(pseudo population) U를 생성한다. 생성된 의사모집단을 각 소지역별로 구별하여 U_j라 하고 그 크기 N_j를 구한다.

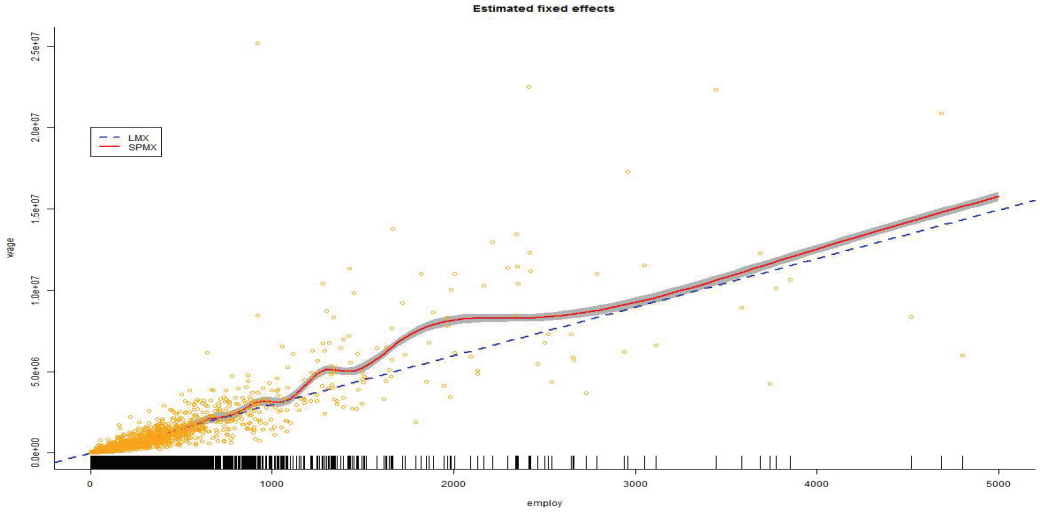

Figure 4.1. Comparison of linear mixed and semiparametric model

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

(3) 의사모집단 U로부터 각 소지역을 층으로 하는 층화추출을 통해 원자료와 크기가 같은 $n = 7{,}038$의 표본을 얻는다. 층화추출 된 표본 내 각 층의 크기가 원자료와 동일한 n_j가 되도록 한다. 각 층에서 추출된 자료를 모은 것을 s_j라 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 $\hat{\bar{Y}}_j$들을 구한다.

(5) 축소소지역추정량을 구하기 위해 (3)에서 추출된 표본을 복원추출로 랜덤하게 $B = 200$번 추출하여 분산을 구하고, 구해진 분산과 임금 추정치를 이용하여 CV를 추정한 후 최종적인 축소소지역추정량을 구한다.

(6) 구해진 각 임금추정치 $\hat{\bar{Y}}_j$와 (2)의 $\bar{Y}_j$을 정해진 비교통계량을 이용하여 비교한다.

(7) (3)–(5)를 $R = 500$회 반복 실시한다.

아래 Figure 4.1은 준모수와 선형혼합모형을 이용한 소지역 추정량을 그림으로 나타낸 것이다. 특히 종사자 수가 2,000명 부근의 그림을 살펴보면 선형이 아닌 비선형 관계가 있는 것으로 나타나 준모수 모형을 적합하는 것이 타당하다고 판단된다. 그러나 Figure 4.2와 Figure 4.3은 선형혼합모형과 축소선형혼합모형, 그리고 준모수혼합모형과 축소준모수혼합모형을 그림으로 나타낸 것인데 서로 구별하기 어려울 정도로 유사한 값으로 예측되었다. 이는 CV가 '0'에 가까운 경우에 나타나는 현상으로 결국 축소 모형의 효과가 크게 없을 것으로 판단된다.

또한 Figure 4.4는 각 모형에서 얻어진 잔차 그림이다. 선형혼합모형에 비해 준모수혼합모형의 잔차 크기가 작은 것을 확인할 수 있으며 특징적으로 준모수혼합모형의 잔차는 점점 커지는 형태를 취하고 있다. 축소선형혼합모형과 축소준모수혼합모형의 잔차는 음수쪽으로 치우쳐져 있는 것을 확인할 수 있는데 이는 축소 추정량을 사용한 결과이므로 자연스러운 결과이다.

다음의 Table 4.1은 500회의 반복실험을 통해 얻은 각종 비교통계량의 값들을 정리한 것이다. 축소추정량을 얻기 위한 붓스트랩 반복수는 200번이다. 결과를 살펴보면 선형혼합모형에 비해 준모수혼합모형

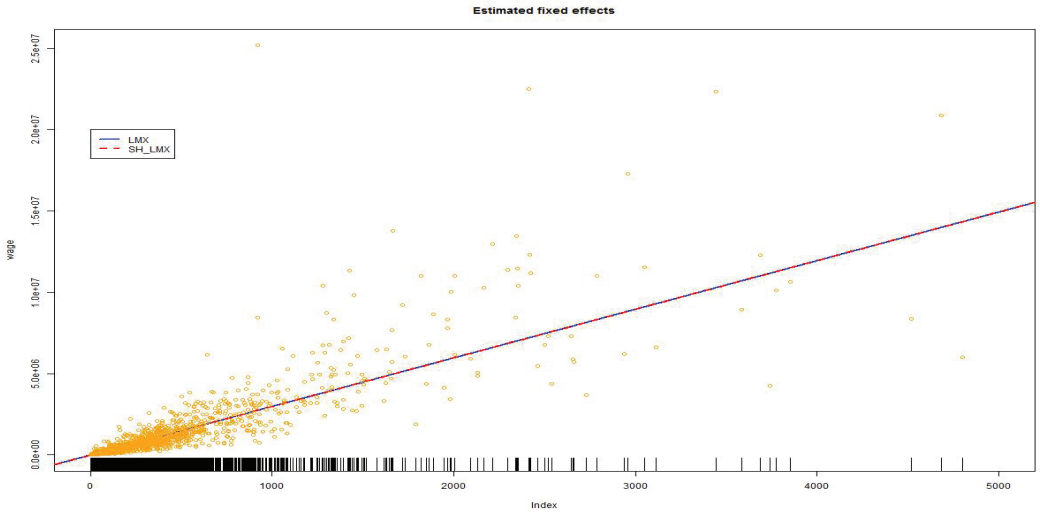

Figure 4.2. Comparison of linear mixed and shrinkage linear mixed model

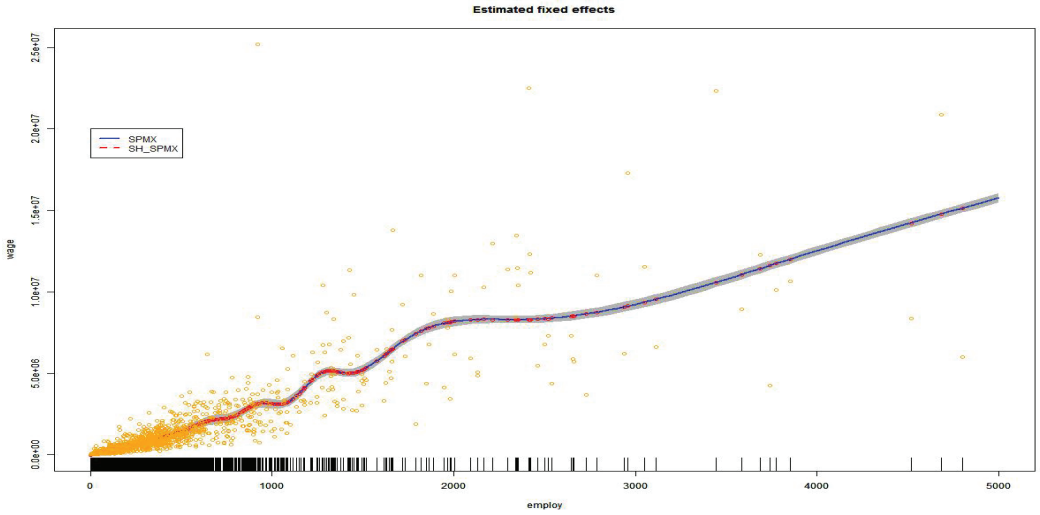

Figure 4.3. Comparison of semiparametric and shrinkage semiparametric model

의 결과가 매우 우수한 것을 확인할 수 있다. 또한 선형혼합모형의 경우에는 축소모형을 적용하여 얻은 축소선형혼합모형을 이용한 추정량의 정확도가 미세하지만 향상된 것을 볼 수 있다. 그러나 축소준모수 혼합모형을 사용할 경우는 준모수혼합모형을 사용한 결과와 유사하지만 미세하게 나쁜 결과를 주고 있다. 따라서 본 자료의 경우에는 준모수혼합모형을 사용하는 것이 타당하다고 판단된다.

5. 결론 및 전망

소지역 추정 방법으로 모형 기반 추정법에 대한 관심이 증대되는 가운데 준모수혼합모형을 이용한 함수 추정 기법은 이 분야에서 활용도가 매우 높을 것으로 기대된다. 또한 MSE를 기준으로 만들어진 추정

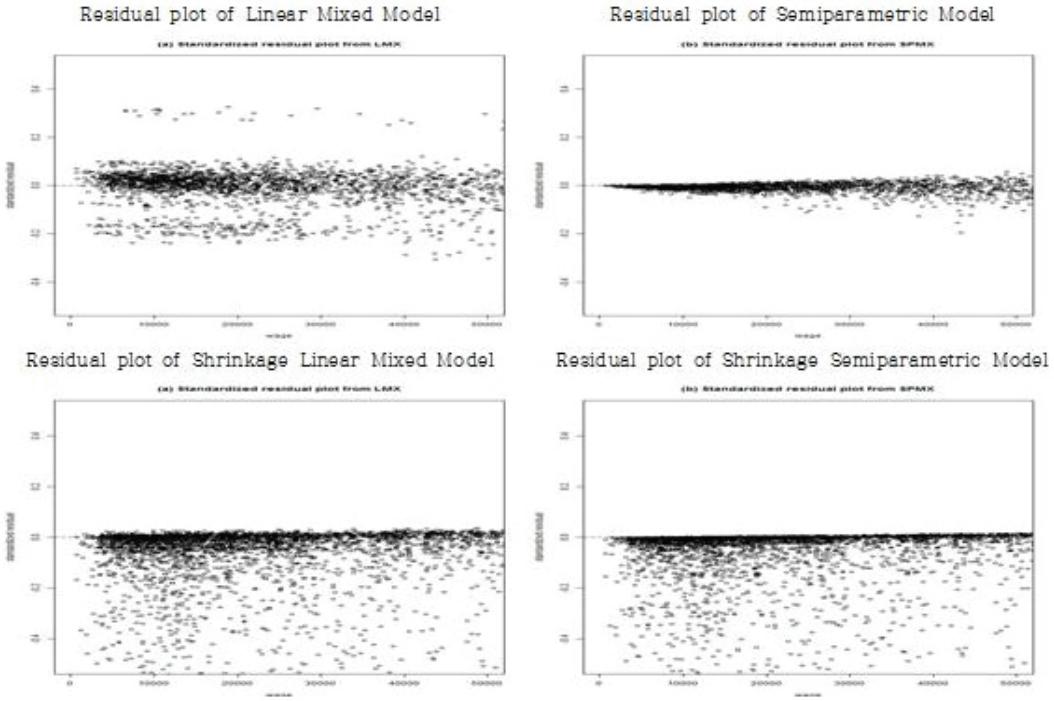

Figure 4.4. Residual comparison of suggested methods

Table 4.1. Comparison results of real data analysis

추정량	MSE($\times 10^8$)	MAE($\times 10^4$)	RE($\times 10^{-2}$)	ARE($\times 10^{-2}$)
선형혼합 $\hat{\bar{Y}}_j^{MX}$	18.2945	2.6898	5.9494	17.6056
	(0.0462)	(0.0033)	(0.0130)	(0.0227)
준모수혼합 $\hat{\bar{Y}}_j^{SPMX}$	1.6644	0.8358	0.6759	5.6443
	(0.0066)	(0.0018)	(0.0026)	(0.0232)
축소선형혼합 $\hat{\bar{Y}}_j^{SH\text{-}MX}$	18.2778	2.6821	5.9215	17.5272
	(0.0464)	(0.0033)	(0.0130)	(0.0226)
축소준모수혼합 $\hat{\bar{Y}}_j^{SH\text{-}SPMX}$	1.6736	0.8385	0.6805	5.6651
	(0.0067)	(0.0019)	(0.0026)	(0.0122)

량에 추가하여 상대오차를 기준으로 만들어진 새로운 추정량은 자료의 형태에 따라 타당한 추정량을 사용할 수 있는 여건을 제공하였다. 모의실험 결과 선형혼합모형과 준모수혼합모형은 모집단 자료 형태가 선형과 비례형인 경우에는 비슷한 결과를 주는 것으로 판단된다. 그러나 볼록형 또는 오목형인 경우에는 준모수혼합모형을 반드시 사용하여야 우수한 결과를 얻을 수 있다. 또한 볼록형 또는 오목형에서는 상대오차를 최소로 하여 만들어진 축소준모수혼합모형을 사용한다면 MSE 또는 MAE 기준도 어느 정도 만족하면서도 상대오차를 크게 줄일 수 있기 때문에 축소준모수혼합모형을 사용할 필요가 있다. 사례연구 결과에서는 준모수혼합모형이 축소준모수혼합모형에 비해 미세하지만 우수한 결과를 주고 있다. 따라서 실제 자료 분석에서는 모집단 자료 형태를 선행적으로 연구한 후, 그 결과에 따라 사용 유무를 결정할 필요가 있다.

References

Hwang, H.-J. and Shin, K.-I. (2008). Shrinkage prediction for small area estimation, *The Korean Journal of Applied Statistics*, **21**, 109–123.

Jeong, S.-O. and Shin, K.-I. (2008). A new nonparametric method for prediction based on mean squared relative error, *Communications of the Korean Statistical Society*, **15**, 255–264.

Jeong, S.-O. and Shin, K.-I. (2013). Semiparametric and nonparametric mixed effects models for small area estimation, *The Korean Journal of Applied Statistics*, **26**, 71–79.

Jones M. C., Park, H., Shin, K.-I., Vines, S. K. and Jeong, S.-O. (2008). Relative error prediction via kernel regression smoother, *Journal of Statistical Planning and Inference*, **138**, 2887–2898.

Lee, H., Rancourt, E. and Sarndal, C.-E. (1995). Experiment with variance estimation from survey data with imputed value, *Journal of Official Statistics*, **10**, 231–243.

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.

Park, H. and Stefanski, L. A. (1997). Relative error prediction, *Statistics and Probability Letters*, **40**, 227–236.

Rao, J. N. K. (2003). *Small Area Estimation*, John Wiley and sons, New York.

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.

준모수혼합모형을 이용한 축소소지역추정

정석오[a] · 추만호[a] · 신기일[a,1]

[a]한국외국어대학교 통계학과

(2014년 3월 20일 접수, 2014년 6월 10일 수정, 2014년 6월 18일 채택)

요 약

소지역추정은 작은 규모의 지역 또는 도메인에 작은 크기의 표본이 배정되어 추정의 정도가 좋지 않은 경우에 이를 극복하는 통계적 기법이다. 소지역추정에 흔히 사용되고 있는 모형기반 추정량은 MSE를 기초로 얻어지나 최근 상대오차를 이용한 소지역추정법도 연구되고 있다. 본 논문에서는 상대오차를 최소로 하는 소지역 추정량의 준모수적 접근법에 관하여 연구하였다. 즉 준모수혼합모형을 이용한 축소소지역추정량을 새롭게 제안하였다. 또한 Lee (1995)에서 제안된 모의실험 자료를 이용한 모의실험과 매월노동통계 자료를 이용한 사례연구를 통하여 기존의 추정량과 제안된 추정량의 우수성을 비교하였다.

주요용어: 절사선형스플라인 모형, 선형혼합모형, 경험적최량불편추정법, 축소추정법.

이 논문은 2014도 한국외국어대학교 교내연구비 지원을 받아 수행되었음.

[1]교신저자: (449-791) 경기도 용인시 처인구 모현면 외대로 81, 한국외국어대학교 통계학과.
E-mail: keyshin@hufs.ac.kr