응용통계연구 제 **21**권 *1*호, *2008*년, *pp. 109–123*

축소예측을 이용한 소지역 추정*

황희진[1] 신기일[2]

요 약

많은 소지역 추정량이 제안되었으며, 국내외에서 소지역 추정에 관한 많은 연구가 진행되고 있다. 또한 소지역 추정량의 특성과 우수성을 비교하기위한 비교통계량도 연구되고 있다. 기존의 소지역 추정량은 MSE(Mean square error)를 최소화하여 얻어지며, 이에 따라 추정량의 우수성도 MSE를 기준으로 판단하고 있다. 본 논문에서는 최근 새롭게 재조명 되고 있는 MSPE(Mean square percentage error)를 최소화하는 추정량을 제안하였다. 신기일 등 (2007)에서 사용된 비교통계량과 MSE 그리고 MSPE를 이용하여 제안된 추정량과 기존의 소지역 추정량을 비교하였다.

주요용어: 소지역통계, 공간통계, 축소예측량(Shrinkage predictor), MSPE(Mean square percentage error).

1. 서론

대부분의 표본조사 자료는 전국 규모의 총계나 평균을 추정하기 위한 것으로 설계되어, 표본설계 시 고려되지 않았던 보다 작은 단위의 지역에 대한 의미있는 추정값을 구하는 것은 쉬운 일이 아니다. 그러나 많은 지방자치단체 등은 의미있는 소지역 추정값을 원하고 있으며 이에 맞는 통계적 추정기법이 필요하게 된다. 이렇게 표본설계 시 계획되지 않은 지역 또는 영역에 관해 추정하고자 할 때 이용되는 통계적 추정방법이 소지역 추정이다 (Rao, 2003). 많은 통계 선진국은 이미 소지역 추정에 관한 연구 결과를 축적하고 있으며, 최근 국내에서도 김달호와 김남희 (2002), Kim과 Choi (2004)와 같이 소지역 추정에 관한 연구가 활발히 진행되고 있다.

여러 소지역 추정량이 발표되었지만 다른 통계 분야와 마찬가지로 소지역 추정법에서도 MSE(Mean square error)를 최소화하는 추정량을 사용하고 있다. 따라서 잘 알려진 것처럼, 이러한 소지역 추정량은 각 소지역의 크기에 상관없이 같은 크기의 오차 또는 등분산성을 전제로 하는 것이며 일반적으로 큰 무리없이 사용되고 있다.

이제 한국의 대표적인 소지역이라 할 수 있는 시.군.구를 고려해보자. 일반적으로 이들의 크기는 매우 큰 차이를 보이고 있다. 서울 강남구의 인구와 지방 작은 군의 인구를 비교

* 이 논문은 2007년도 한국외국어대학교 학술연구비 지원에 의해 수행되었음

1) (449-791) 경기도 용인시 처인구 모현면 왕산리 산 89, 한국외국어대학교 통계학과, 박사과정.
E-mail: lshhj01@naver.com

2) (449-791) 교신저자. 경기도 용인시 처인구 모현면 왕산리 산 89, 한국외국어대학교 정보통계학과, 교수.
E-mail: keyshin@hufs.ac.kr

하면 그 크기의 차이를 알 수 있다. 이렇게 큰 크기의 차이가 있을 때, 같은 크기의 오차를 가정하여 얻어진 추정의 결과를 설명하기에는 때로 어려움이 있을 수 있다. 예를 들어 인구 10만 명인 A 소지역과 인구 50만 명인 B 지역의 실업자를 추정한다고 가정하자. A 지역의 실업자 수는 5,000명, B 지역의 실업자 수는 25,000명이라 하자. 그리고 등분산성을 가정하여 실제 오차는 각 4,000명이라 하자. 그러면 A 지역의 오차 비율은 80%가 될 것이고 B 지역의 오차 비율은 16%가 될 것이다. 오차 비율 80%를 주는 통계는 A 지역에 사는 시민들에게 있어 납득할 만한 수준의 통계라고 말하기가 어려울 것이다. 이러한 현상은 MSE를 기준으로 했을 때 언제나 발생할 가능성이 있다. 일반적으로 모든 자료에서 오차 비율이 작다면 설명에 큰 어려움이 없을 것이다. 그러나 크기가 작은 지역에서 오차 비율이 크게 나타나 해석에 어려움이 있다면 이를 극복하기 위한 방법이 필요하며, MSPE가 대안이 될 수 있을 것이다.

물론 등분산성 가정은 모든 소지역 추정법에 적용되지 않는다. 등분산성 가정은 설계기반 추정인 직접, 합성 그리고 복합 추정법에서 사용된다. 또한 모형기반 추정법인 회귀분석 추정법, 공간 추정법 등에서 사용된다. 최근 소지역 추정법으로 우수한 결과를 주고 있는 경험적 베이즈 추정법(Empirical Bayese estimation)과 계층적 베이즈 추정법(Hierarchical Bayese estimation) 등은 이러한 가정을 하지 않고 있으며 또한 Fay-Herriot 모형 (Fay와 Herriot, 1979)은 각 지역의 이분산성도 가정하고 있다. 따라서 이들 추정량과 MSPE를 바탕으로 제안된 추정량을 적절히 융합하는 새로운 방법도 연구할 필요가 있다.

본 논문에서는 MSPE(Mean square percentage error)를 기준으로 하여 얻어지는 추정량을 제안하였다. 또한 신기일 등 (2007)에서 사용했던 비교통계량을 기준으로 MSPE를 이용한 새로운 소지역 추정량과 기존의 소지역 추정량을 비교하였으며 MSE와 MSPE를 이용한 비교도 함께 실시하였다.

본 논문의 구성은 다음과 같다. 2절에서는 제안된 기존의 추정량과 MSPE를 기준으로 하여 얻어지는 추정량에 대하여 살펴보았고, 3절에서는 각 소지역 추정량들을 비교할 수 있는 여러 가지 비교통계량을 소개하였다. 4절에서는 실제 자료를 이용하여 본 논문에서 제안한 추정량과 기존의 추정량을 3절에서 살펴본 비교 통계량을 사용하여 비교하였으며 또한 모의실험을 통하여 기존에 제안되었던 추정량들과 비교분석하였다. 이때 사용된 비교통계량은 전통적인 비교 통계량인 MSE와 MSPE이다. 끝으로 5절에 결론이 있다.

2. 소지역 추정량

2.1. MSPE를 이용한 소지역 추정량

2.1.1. 축소 예측량(Shrinkage predictor : $\hat{Y}^{SH}$)

기존의 추정량을 구하는 기준은 MSE(Mean square error), 즉, $E(Y-\hat{Y})^2$를 최소화하는 것이었다. 이 기준은 Y값에 상관없이 같은 크기의 오차를 가정한 모형이 적당할 때 사용하면 우수한 결과를 준다. 그러나 큰 Y값에서는 상대적으로 오차의 크기가 문제 되지 않지만, 작은 Y값에 대해서는 상대적으로 오차가 매우 클 수 있으며 이 경우 MSE를 기준으로 추정량을 구하고 또한 추정량의 우수성을 파악하는 것 보다는 MSPE를 사용하는 것이 타

당하다. 따라서 이에 알맞는, 즉, MSE 대신 MSPE(Mean square percentage error)를 기준으로 하여 구해지는 추정량이 필요하게 된다. MSPE를 이용하여 소지역 추정량을 구하는 방법을 살펴보자. MSPE를 이용한 예측량은

$$\min E\left(\frac{Y-\hat{Y}}{Y}\right)^2$$

를 만족하는 $\hat{Y}$로 구해지며 Park과 Stefanski (1997)에 의하면 $\hat{Y}$는 다음과 같이 얻어진다.

$$\hat{Y}=\frac{E\left(\frac{1}{Y}\right)}{E\left(\frac{1}{Y^2}\right)}. \tag{2.1}$$

(2.1)식을 사용하는데 있어서의 문제점은 먼저 $E(1/Y)$와 $E(1/Y^2)$이 존재해야 하며, 또한 각각의 기댓값을 구하는 것이 분포에 따라서는 쉽지 않다는 것이다. 그리고 자료가 모두 "0" 보다 크다는 조건이 전제되어야 한다. 이러한 문제를 해결하기 위하여 본 논문에서는 먼저 평균과 분산에 대하여 다음을 가정하였다.

$$\left(\frac{Y-\mu}{\mu}\right)^m=o_p(1), \qquad m=2,3,\ldots,\mu=E(Y).$$

그러면 Taylor 전개에 의해

$$\begin{aligned}\frac{1}{Y} &= \frac{1}{\mu}-\frac{1}{\mu^2}(Y-\mu)+\frac{1}{\mu^2}(Y-\mu)^2+o_p(1),\\ \frac{1}{Y^2} &= \frac{1}{\mu^2}-\frac{2}{\mu^3}(Y-\mu)+\frac{3}{\mu^4}(Y-\mu)^2+o_p(1)\end{aligned}$$

이 되고 이를 이용하여 다음의 결과를 얻는다.

$$\begin{aligned}E\left(\frac{1}{Y}\right) &\approx \mu^{-1}(1+CV^2),\\ E\left(\frac{1}{Y^2}\right) &\approx \mu^{-2}(1+3CV^2).\end{aligned} \tag{2.2}$$

여기서 $CV=\sigma/\mu$이다. 이제 (2.2)식을 (2.1)식에 대입하면

$$\frac{E\left(\frac{1}{Y}\right)}{E\left(\frac{1}{Y^2}\right)}\approx\frac{\mu^{-1}(1+CV^2)}{\mu^{-2}(1+3CV^2)}=\mu\frac{(1+CV^2)}{(1+3CV^2)}\approx\mu(1-2CV^2) \tag{2.3}$$

이 된다. 이제 CV가 작다고 가정하고 μ대신 $\hat{Y}$를 대입하면 다음의 예측량을 얻는다.

$$\hat{Y}^{SH}=\hat{Y}(1-2CV^2). \tag{2.4}$$

위의 식에서 보듯이 MSE기준 예측량보다 예측값이 작아지는 것을 확인할 수 있다. 일반적으로 MSPE 기준 예측량은 축소예측량(Shrinkage predictor)이 되므로 본 논문에서도 축소예측량이라 부르겠다. 위의 축소예측량의 가장 큰 특징은 기존의 추정량에 간단히 $(1-2CV^2)$를 곱하여 얻을 수 있다는 것이다.

만약 CV가 충분히 작지 않아 (2.4)식의 값이 음수 값을 갖게 된다면

$$\hat{Y}^{SH} = \hat{Y}\frac{(1+CV^2)}{(1+3CV^2)} \tag{2.5}$$

또는

$$\hat{Y}^{SH} = \hat{Y}\exp\left(-2CV^2\right) \tag{2.6}$$

을 사용하면 된다. 본 논문에서는 큰 CV값으로 인하여 몇몇 소지역에서 음수가 발생하여 (2.5)식과 (2.6)식을 사용하였다.

2.2. 기존의 소지역 추정량

많은 소지역 추정량이 제안되었으나 본 논문에서는 신기일 등 (2007)에서 연구된 소지역 추정량의 일부를 중심으로 살펴보았다. 전술하였듯이 어떤 소지역 추정량이 구해지면 간단히 수정하여 축소 소지역 추정량으로 사용할 수 있고 또한 본 논문의 목적이 축소 소지역 추정량의 특징을 파악하는 것이므로 일부 추정량만을 선택하여 분석하였다.

2.2.1. 자료기반추정량

자료기반 소지역 추정법으로는 직접추정법(Direct estimation), 합성추정법(Synthetic estimation), 그리고 복합추정법(Composite estimation)이 있다. 직접추정법은 해당 소지역에 배정된 표본만을 이용하여 추정하는 방법이며 합성추정법은 추정하고자 하는 소지역과 특성이 유사한 다른 소지역들의 정보를 이용하여 추정값의 정도를 높이고자 하는 추정법이다. 복합추정법은 직접추정량의 큰 분산과 합성추정량의 편향을 보완하기 위해서 직접추정량과 합성추정량의 가중평균을 사용하는 방법이다. 이 방법들 중 본 논문에서는 직접추정량만을 고려하였다. 직접추정량은 소지역 추정시 해당 지역에 해당되는 자료만을 이용하여 추정하는 방법으로 비교적 간단히 구할 수 있으며 불편성은 만족하나 표본수가 불균형적일 뿐만 아니라 자료수가 적어서 추정오차가 크다는 단점이 있어 이용에는 어려움이 있다. 본 논문에서 사용된 직접추정량($\hat{Y}_{DE}$)은 다음과 같이 정의된다.

$$\hat{Y}_{DE} = \hat{Y}_i = \sum_j w_{ij}y_{ij}.$$

여기서 $\hat{Y}_i$는 i번째 소지역 값을 의미하고 w_{ij}는 가중치를, y_{ij}는 i지역 j번째 자료값을 나타낸다.

2.2.2. 모형기반추정법

모형기반추정법으로는 회귀분석방법, 경험적 베이즈 추정법(Empirical Bayese estimation), 계층적 베이즈 추정법(Hierarchical Bayese estimation) 등이 있으며 일반적으로 계층적베이지안 추정법이 우수한 것으로 알려져 있다. 이중 가장 많이 이용되고 있는 것은 회귀모형을 이용한 추정 방법이다. 일반적인 회귀모형에서 회귀계수를 추정하고 이를 이용하여 소지역 추정량으로 회귀모형의 예측통계량을 이용하는 것으로 회귀추정량($\hat{Y}_{REG}$)은 다음과 같다.

$$\hat{Y}_{REG} = \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki}.$$

여기서 $\hat{Y}_i$는 i번째 소지역 추정값이며 $x_{1i}, \ldots, x_{ki}$는 이에 해당되는 설명변수이고 $\hat{\beta}_i$는 추정된 회귀계수이다.

2.2.3. 공간추정량($\hat{Y}_{SP}$)

일반적으로 모형기반(model based) 추정량은 자료기반 추정량에 비해 우수한 결과를 주는 것으로 알려져 있다. 그러나 모형기반 추정량은 충분한 설명변수가 있다는 가정 하에 얻어지는 추정량이며 이 가정이 충족되지 못할 경우에는 그 사용이 제한적일 수밖에 없다. Shin과 Lee (2003)은 공간통계학을 이용한 소지역 추정법을 제안하였으며 이 논문에서 설명변수로 관심변수의 공간상관관계를 이용할 경우 더 좋은 결과를 얻을 수 있음을 보였다. 이때 얻어진 추정량을 공간추정량이라 하였으며 공간추정량 $\hat{Y}_{SP}$는 다음의 식에서 구해진다.

$$\hat{Y}_{SP} = \hat{Y}_i = \hat{\rho} S_i.$$

여기서 $\hat{Y}_i$는 i번째 소지역 추정값이며 S_i는 i번째 소지역의 이웃을 모두 더하여 얻어진 공간 변수이고, $\hat{\rho}$는 추정된 계수이다. 이에 관한 자세한 사항은 Kim과 Shin (2006)에 나와 있다.

2.2.4. 선형결합추정량

선형결합 소지역 추정량은 일반적으로 불편추정량인 직접 추정량과 분산을 줄일 수 있는 모형기반 추정량의 선형결합으로 이루어진다. 예를 들면 직접추정량과 공간추정량을 선형결합함으로써 직접 추정량의 변동성과 공간추정량의 편향을 동시에 줄일 수 있게 된다. 직접추정량과 공간추정량의 선형결합추정량은 다음과 같다.

$$\hat{Y}_{DESP} = \alpha_{SP} \hat{Y}_{DE} + (1 - \alpha_{SP}) \hat{Y}_{SP}.$$

여기서 가중치 α_{SP}는 $\hat{Y}_{SP}$와 $\hat{Y}_{DE}$의 MSE를 사용하여야 하나 많은 경우 MSE를 사용하는 대신 각 추정량의 분산을 이용하여 구한다. 즉

$$\alpha_{SP} = \frac{\mathrm{Var}(\hat{Y}_{SP})}{\mathrm{Var}(\hat{Y}_{DE}) + \mathrm{Var}(\hat{Y}_{SP})}$$

을 이용하여 구한다. 이에 관한 자세한 설명은 Rao (2003) 또는 신기일 등 (2007)을 살펴보기 바란다. 본 논문에서는 선형결합추정량으로 $\hat{Y}_{SP}$와 $\hat{Y}_{REG}$를 결합한 $\hat{Y}_{SPREG}$, $\hat{Y}_{DE}$와 $\hat{Y}_{REG}$를 결합한 $\hat{Y}_{DEREG}$, 그리고 $\hat{Y}_{DE}$와 $\hat{Y}_{SPREG}$를 결합한 $\hat{Y}_{DESPREG}$를 사용하였다.

2.3. 비교된 소지역 추정량

(2.2)절에서 소개된 소지역 추정량은 다음과 같다.

$$\hat{Y}_{DE},\ \hat{Y}_{SP},\ \hat{Y}_{REG},\ \hat{Y}_{SPREG},\ \hat{Y}_{DESP},\ \hat{Y}_{DEREG},\ \hat{Y}_{DESPREG}.$$

다음으로 이 추정량을 이용하여 만들어진 축소 소지역 추정량은 (2.5)식과 (2.6)식으로 다음과 같다.

$$\hat{Y}^{SH} = \hat{Y}\frac{(1+CV^2)}{(1+3CV^2)} \tag{2.7}$$

또는

$$\hat{Y}^{SH} = \hat{Y}\exp(-2CV^2). \tag{2.8}$$

따라서 본 논문에서 사용할 축소 소지역 추정량은 다음과 같다.

$$\hat{Y}_{DE}^{SH} = \hat{Y}_{DE}\frac{(1+CV^2)}{(1+3CV^2)} \quad \text{and} \quad \hat{Y}_{DE}^{SH} = \hat{Y}_{DE}\exp(-2CV^2),$$

$$\hat{Y}_{SP}^{SH} = \hat{Y}_{SP}\frac{(1+CV^2)}{(1+3CV^2)} \quad \text{and} \quad \hat{Y}_{SP}^{SH} = \hat{Y}_{SP}\exp(-2CV^2),$$

$$\hat{Y}_{REG}^{SH} = \hat{Y}_{REG}\frac{(1+CV^2)}{(1+3CV^2)} \quad \text{and} \quad \hat{Y}_{REG}^{SH} = \hat{Y}_{REG}\exp(-2CV^2),$$

$$\hat{Y}_{SPREG}^{SH} = \hat{Y}_{SPREG}\frac{(1+CV^2)}{(1+3CV^2)} \quad \text{and} \quad \hat{Y}_{SPREG}^{SH} = \hat{Y}_{SPREG}\exp(-2CV^2),$$

$$\hat{Y}_{DESP}^{SH} = \hat{Y}_{DESP}\frac{(1+CV^2)}{(1+3CV^2)} \quad \text{and} \quad \hat{Y}_{DESP}^{SH} = \hat{Y}_{DESP}\exp(-2CV^2),$$

$$\hat{Y}_{DEREG}^{SH} = \hat{Y}_{DEREG}\frac{(1+CV^2)}{(1+3CV^2)} \quad \text{and} \quad \hat{Y}_{DEREG}^{SH} = \hat{Y}_{DEREG}\exp(-2CV^2),$$

$$\hat{Y}_{DESPREG}^{SH} = \hat{Y}_{DESPREG}\frac{(1+CV^2)}{(1+3CV^2)} \quad \text{and} \quad \hat{Y}_{DESPREG}^{SH} = \hat{Y}_{DESPREG}\exp(-2CV^2).$$

3. 비교통계량

소지역 추정량을 평가하기 위한 여러 비교통계량이 제안되었으며 그 중 본 논문에서 사용된 통계량은 R^2와 기울기, 커버리지, 캘리브레이션 그리고 MSE와 MSPE이다. 이에 관한 자세한 내용은 신기일 등 (2007)을 살펴보기 바라며 본 논문에서는 이를 간단히 설명하였다.

3.1. 회귀모형을 이용한 방법(R^2과 기울기)

회귀모형을 이용한 진단방법은 직접 추정량이 불편 추정량임을 활용하여 비교대상 추정량의 불편성을 진단하는 것으로 내용은 다음과 같다. 직접추정량을 종속변수로 하고 비교대상 추정량을 독립변수로 하는 절편이 없는 단순회귀모형을 만들고 단순회귀식을 적합한 후 이 때 얻어지는 결정계수 R^2값과 기울기를 비교해 본다. 만약 기울기가 "1"에서 많이 떨어져 있거나 R^2값이 "1"보다 많이 작다면 좋은 추정량이라고 할 수 없다.

3.2. 커버리지(Coverage)

커버리지는 직접추정량의 95% 신뢰구간을 구하고 이 구간에 각각의 추정량에서 얻어진 추정치의 몇 퍼센트가 포함되는가를 살펴보는 것으로 모형기반추정량이 편향이 없고 분산이 작다면 대부분 이 구간에 포함되어 높은 커버리지를 보일 것이다. 잘 알려진 것처럼 직접추정량의 분산은 매우 크며 따라서 95% 신뢰구간 또한 매우 크다. 따라서 비교 통계량은 100% 커버리지를 보이는 것이 일반적이어야 한다. 만약 커버리지가 100%가 아닌, 작은 값을 갖는다면 이는 비교 추정량의 편향이 크다든지 하는 등의 문제가 있다는 증거일 수 있다. 즉 본 논문에서 사용된 커버리지는 편향의 크기를 알 수 있는 비교 통계량이다.

3.3. 캘리브레이션(Calibration)

여러 개의 소지역을 합치게 되면 그 지역에 포함된 표본의 수가 증가하게 되고 따라서 일반적으로 직접추정량의 정도가 높아지게 된다. 이러한 특징을 이용하여 여러 소지역을 합쳐가면서 각 추정량에서 얻어진 추정치들을 직접추정량과 비교하는 것을 캘리브레이션이라고 한다. 만약 지역이 커져 자료의 수가 늘어났음에도 불구하고 직접추정치와 모형기반추정치가 큰 차이를 보인다면 모형기반추정량이 자료를 잘 설명한다고 판단할 수 없다.

3.4. MSE와 MSPE

MSE와 MSPE는 추정량 비교에 일반적으로 사용되는 비교 통계량이다. MSE와 MSPE의 정의는 다음과 같다.

$$\begin{aligned} \text{MSE} &= \frac{1}{R}\frac{1}{n}\sum_{r=1}^{R}\sum_{i=1}^{n}(Y_i - \hat{Y}_i^{(r)})^2, \\ \text{MSPE} &= \frac{1}{R}\frac{1}{n}\sum_{r=1}^{R}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{Y}_i^{(r)}}{Y_i}\right)^2. \end{aligned}$$

여기서 $i = 1, \ldots, n$은 i번째 소지역을 의미하며 본 논문에서는 36개의 소지역이 있으므로 $n = 36$이 된다. 모의실험에서 사용된 반복수 $R = 5,000$이 사용되었다. 이제 MSE, MSPE를 구하기 위해서는 참값이 필요하게 된다. 그러나 참값을 알 수 없기 때문에 본 논문에서는 Kim과 Shin (2006)에서 사용했던 방법을 고려하였다. 즉 9,000여개의 자료에서 얻어진 $\hat{Y}_{DE}$를 참값 Y_i라 가정하였다. 다음으로 9,000개의 자료에서 3,000, 4,000 그리고 5,000개의

자료를 랜덤 추출한 후 각 소지역 추정량을 이용하여 계산한 결과를 $\hat{Y}_i^{(r)}$이라 하였다. 이렇게 얻어진 값을 R번 반복한 후 평균을 구하게 되면 36개 소지역별로 MSE_i와 MSPE_i를 얻게 되며 최종적으로 이를 평균한 값을 표에 작성하였다.

4. 자료분석 및 추정량 비교

4.1. 자료분석

본 논문에서 사용한 자료는 2005년 실업자수를 이용하였으며 이 자료는 신기일 등 (2007)의 논문에서 사용된 자료와 같다. 그러나 실업자수가 "0"인 일부 시군구자료가 포함되어 있어, 본 논문의 핵심인 MSPE를 비교하기위해 이 자료를 분석에서 제외하였다. 이로 인해 신기일 등 (2007)의 결과와 $\hat{Y}_{DE}$등 이와 관련이 없는 경우는 소지역 추정량이 완전히 일치하나 다른 소지역 추정량은 약간 결과가 다르게 나왔다. 또한 선형결합추정량을 구하기 위한 가중치 α_*값은 Bootstrap 방법을 이용하여 구하였다. 2005년도 실업자 자료 9,000여개를 복원추출하여 붓스트랩 샘플을 생성하고 5,000번 반복하여 분산을 구하였으며 (2.2.4)절의 공식을 이용하여 가중치 α_*값을 추정하였다.

4.2. 추정량 비교

4.2.1. 기울기와 R^2

2.3절에서 제안된 20개의 소지역 추정량이 비교 되었다. 먼저 $\hat{Y}_{DE}$를 종속변수로, 각 소지역 추정량을 독립변수로 한 R^2과 기울기를 구하였다. 결과는 표 4.1에 나타냈다. 먼저 기존의 소지역 추정량의 기울기와 R^2을 살펴보자. 기울기 "1"에서 가장 많이 벗어난 추정량은 $\hat{Y}_{DEREG}$이고 그 다음으로 $\hat{Y}_{SP}, \hat{Y}_{SPREG}$등이다. 그러나 크게 "1"에서 벗어나지 않아 큰 문제는 없다고 할 수 있다. 결정계수 R^2을 살펴보면 $\hat{Y}_{REG}$가 약 0.80 정도이고 다음이 $\hat{Y}_{DEREG}$이 0.85로 낮은 값을 갖고 있다. 이에 비해 축소 소지역 추정량들의 기울기는 모두 크게 "1"에서 벗어난 것을 확인할 수 있다. 이는 독립변수에 해당하는 소지역 추정량들이 모두 축소 추정량이 되었기 때문에 "x"축이 축소가 되어 모든 기울기가 "1"을 크게 벗어나고 있다. 그러나 $\hat{Y}_{REG}^{SH}$는 "1"을 크게 벗어나지 않고 있다. 이는 축소가 다른 추정량에 비해 작게 일어났기 때문으로 풀이된다. 다음으로 결정계수 R^2을 살펴보면 큰 변화가 없음을 확인 할 수 있다. 즉, $\hat{Y}_{REG}$가 약 0.80 정도이고 다음이 $\hat{Y}_{DEREG}$이 0.86 정도로 낮은 값을 갖고 있으며 그 차이도 매우 미미하다. 결론적으로 축소 소지역 추정량은 불편 추정량이 아닌 것을 다시 한번 확인하였다. 이에 반하여 결정계수는 큰 차이를 보이지 않는다.

표 4.1: 추정량별 기울기와 R^2

	추정량	기울기	R^2
$\hat{Y}$	$\hat{Y}_{SP}$	0.9733	0.9749
	$\hat{Y}_{REG}$	1.0000	0.8023
	$\hat{Y}_{SPREG}$	0.9734	0.9763
	$\hat{Y}_{DESP}$	1.0152	0.9927
	$\hat{Y}_{DEREG}$	1.0606	0.8528
	$\hat{Y}_{DESPREG}$	1.0127	0.9932
$\hat{Y}\frac{(1+CV^2)}{(1+3CV^2)}$	$\hat{Y}^{SH}_{DE}$	1.2352	0.9931
	$\hat{Y}^{SH}_{SP}$	1.2199	0.9792
	$\hat{Y}^{SH}_{REG}$	1.0443	0.8028
	$\hat{Y}^{SH}_{SPREG}$	1.2191	0.9821
	$\hat{Y}^{SH}_{DESP}$	1.2430	0.9899
	$\hat{Y}^{SH}_{DEREG}$	1.1322	0.8583
	$\hat{Y}^{SH}_{DESPREG}$	1.2411	0.9920
$\hat{Y}e^{-2CV^2}$	$\hat{Y}^{SH}_{DE}$	1.3042	0.9778
	$\hat{Y}^{SH}_{SP}$	1.3012	0.9709
	$\hat{Y}^{SH}_{REG}$	1.0466	0.8028
	$\hat{Y}^{SH}_{SPREG}$	1.2999	0.9758
	$\hat{Y}^{SH}_{DESP}$	1.3072	0.9799
	$\hat{Y}^{SH}_{DEREG}$	1.1377	0.8588
	$\hat{Y}^{SH}_{DESPREG}$	1.3060	0.9833

4.2.2. 커버리지

$\hat{Y}_{DE}$를 기준으로 커버리지는 구해지며 이를 통하여 편향의 정도를 알 수 있다. 즉, 90% 이하인 커버리지 개수가 많으면 적은 개수를 갖는 추정량에 비해 편향이 크다는 것을 말해준다.

먼저 붓스트랩을 이용하여 $\hat{Y}_{DE}$의 분산을 구한 후 이를 이용하여 95%신뢰구간을 구한다. 구해진 신뢰구간을 이용하여 각 추정량에서 얻어진 추정값이 구해진 $\hat{Y}_{DE}$의 신뢰구간에 얼마나 포함되는지 살펴본다. 이상의 결과를 표 4.2에 나타내었다. 결과를 살펴보면 $\hat{Y}_{DESP}$와 $\hat{Y}_{DESPREG}$가 우수한 것으로 $\hat{Y}_{REG}$와 $\hat{Y}_{DEREG}$가 안 좋은 결과를 주고 있다. 이러한 추세는 축소 소지역 추정량에 그대로 유지된다. 즉, $\hat{Y}^{SH}_{DESP}$와 $\hat{Y}^{SH}_{DESPREG}$가 우수한 결과를 $\hat{Y}^{SH}_{REG}$와 $\hat{Y}^{SH}_{DEREG}$가 안 좋은 결과를 주고 있다. 그러나 전체적으로는 모든 축소추정량에서 향상된 결과를 주고 있다.

표 4.2: 지역별 커버리지

지역	$\hat{Y}_{SP}$	$\hat{Y}_{REG}$	$\hat{Y}_{SPREG}$	$\hat{Y}_{DESP}$	$\hat{Y}_{DEREG}$	$\hat{Y}_{DESPREG}$
1	90	92	91	96	96	96
2	94	100	94	94	100	94
3	100	100	100	99	100	99
⋮	⋮	⋮	⋮	⋮	⋮	⋮
35	93	99	93	94	100	94
36	98	98	99	98	98	98
90%이하인 개수	3	11	2	0	8	0
최소값	75	0	77	90	3	90

$\hat{Y}\frac{(1+CV^2)}{(1+3CV^2)}$으로 계산할 경우

지역	$\hat{Y}^{SH}_{DE}$	$\hat{Y}^{SH}_{SP}$	$\hat{Y}^{SH}_{REG}$	$\hat{Y}^{SH}_{SPREG}$	$\hat{Y}^{SH}_{DESP}$	$\hat{Y}^{SH}_{DEREG}$	$\hat{Y}^{SH}_{DESPREG}$
1	97	96	88	96	97	92	97
2	98	98	100	98	98	100	98
3	98	100	100	100	100	100	100
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
35	100	100	100	100	100	100	100
36	100	100	100	100	100	100	100
90%이하인 개수	0	1	11	1	0	6	0
최소값	95	87	0	87	95	7	95

$\hat{Y}e^{-2CV^2}$으로 계산할 경우

지역	$\hat{Y}^{SH}_{DE}$	$\hat{Y}^{SH}_{SP}$	$\hat{Y}^{SH}_{REG}$	$\hat{Y}^{SH}_{SPREG}$	$\hat{Y}^{SH}_{DESP}$	$\hat{Y}^{SH}_{DEREG}$	$\hat{Y}^{SH}_{DESPREG}$
1	97	96	88	96	96	92	96
2	98	98	100	98	98	100	98
3	98	100	100	100	100	100	100
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
35	100	100	100	100	100	100	100
36	100	100	100	100	100	100	100
90%이하인 개수	0	1	10	1	0	6	0
최소값	95	87	0	87	94	9	94

4.2.3. 캘리브레이션

본 논문에서 사용된 그룹은 신기일 등 (2007)에서 사용한 것과 같은 방법을 사용하였다. 그러나 이중에서 실업자 자료가 "0"인 소지역을 분석에서 제외하였다. 이것은 자료분석에 영향을 주지 않으나 이로 인해 전술한 바와 같이 신기일 등 (2007)의 결과와 약간의 차이를 보이고 있다. 표 4.3을 살펴보면 축소 소지역 추정량의 경우 전체적으로 약 30%정도 작아진 값으로 나타났다. 반면 작은 분산을 보이고 있는 $\hat{Y}_{REG}$와 $\hat{Y}_{DEREG}$는 축소 추정량 값이 상대적으로 작게 줄어들었음을 알 수 있다.

표 4.3: 그룹별 캘리브레이션 결과

GROUP	$\hat{Y}_{DE}$	$\hat{Y}_{SP}$	$\hat{Y}_{REG}$	$\hat{Y}_{SPREG}$	$\hat{Y}_{DESP}$	$\hat{Y}_{DEREG}$	$\hat{Y}_{DESPREG}$
1-1	20962	22563	14348	22168	21519	14359	21353
1-2	15763	15466	18021	15640	15600	16284	15648
1-3	26046	22342	19925	22567	23454	20312	23657
1-4	8933	9762	14633	10077	9509	13768	9691
1-5	27826	24564	13714	23494	25726	14395	25035
1-6	36138	36769	35433	38293	35340	34450	36341
1-7	16322	16600	22726	16595	16888	21451	16846
1-8	21630	24679	19885	24067	22202	18921	22049
1-9	10765	12240	17326	12338	11666	16142	11683
1-10	9753	8355	17335	8500	9008	15775	9068
1-11	1318	1653	2105	1464	1492	1846	1398

$\hat{Y}\frac{(1+CV^2)}{(1+3CV^2)}$으로 계산할 경우

GROUP	$\hat{Y}^{SH}_{DE}$	$\hat{Y}^{SH}_{SP}$	$\hat{Y}^{SH}_{REG}$	$\hat{Y}^{SH}_{SPREG}$	$\hat{Y}^{SH}_{DESP}$	$\hat{Y}^{SH}_{DEREG}$	$\hat{Y}^{SH}_{DESPREG}$
1-1	16711	18137	13891	17801	17363	13928	17224
1-2	11242	11168	17296	11270	11216	15230	11235
1-3	20513	18671	19211	18921	19200	19350	19411
1-4	5345	8009	13976	8256	7392	12817	7496
1-5	23738	21717	13114	20640	22455	13632	21771
1-6	28564	29620	33453	31155	28411	32049	29370
1-7	11759	13299	21654	13290	13400	19940	13333
1-8	17484	18151	18737	17664	17606	17330	17461
1-9	7041	9002	16626	9102	8228	15109	8256
1-10	6317	4866	16690	5008	5574	14817	5636
1-11	789	1132	1788	972	962	1440	887

$\hat{Y}e^{-2CV^2}$으로 계산할 경우

GROUP	$\hat{Y}^{SH}_{DE}$	$\hat{Y}^{SH}_{SP}$	$\hat{Y}^{SH}_{REG}$	$\hat{Y}^{SH}_{SPREG}$	$\hat{Y}^{SH}_{DESP}$	$\hat{Y}^{SH}_{DEREG}$	$\hat{Y}^{SH}_{DESPREG}$
1-1	15559	16970	13875	16656	16299	13901	16171
1-2	9394	9455	17265	9565	9461	15152	9484
1-3	19066	17946	19183	18210	18265	19301	18487
1-4	3223	7408	13944	7655	6533	12746	6613
1-5	22948	21320	13062	20229	21946	13553	21252
1-6	26552	27767	33243	29411	26662	31798	27656
1-7	9967	12230	21572	12266	12402	19780	12341
1-8	16048	15844	18640	15380	16070	17139	15929
1-9	5211	7741	16595	7896	6790	15036	6848
1-10	4571	2757	16663	2936	3697	14744	3776
1-11	480	909	1733	744	697	1329	622

표 4.4: MSE, MSPE를 이용한 비교

	추정량	3000개 표본		4000개 표본		5000개 표본	
		MSE	MSPE	MSE	MSPE	MSE	MSPE
$\hat{Y}$	$\hat{Y}_{DE}$	14,702,227	0.7096	9,086,444	0.4373	5,772,927	0.2754
	$\hat{Y}_{SP}$	19,407,347	0.6927	12,403,498	0.4985	8,202,610	0.4098
	$\hat{Y}_{REG}$	11,025,615	2.3641	10,381,471	2.1625	10,027,633	2.0754
	$\hat{Y}_{SPREG}$	26,313,248	0.8669	12,222,369	0.4521	8,169,461	0.3640
	$\hat{Y}_{DESP}$	13,583,562	0.5828	8,480,360	0.3681	5,477,728	0.2486
	$\hat{Y}_{DEREG}$	8,561,994	0.9570	7,914,967	0.8852	7,553,668	0.8543
	$\hat{Y}_{DESPREG}$	15,767,683	0.6383	8,491,309	0.3644	5,490,689	0.2462
$\hat{Y}\frac{(1+CV^2)}{(1+3CV^2)}$	$\hat{Y}^{SH}_{DE}$	9,975,508	0.3690	6,655,092	0.2595	4,661,897	0.1930
	$\hat{Y}^{SH}_{SP}$	11,801,235	0.4039	7,883,653	0.3263	5,534,652	0.2874
	$\hat{Y}^{SH}_{REG}$	10,768,109	1.8843	10,174,821	1.7449	9,845,150	1.6871
	$\hat{Y}^{SH}_{SPREG}$	15,862,122	0.4960	7,714,535	0.2880	5,394,975	0.2418
	$\hat{Y}^{SH}_{DESP}$	9,771,534	0.3314	6,599,894	0.2339	4,687,659	0.1758
	$\hat{Y}^{SH}_{DEREG}$	8,338,325	0.7242	7,753,991	0.6727	7,423,369	0.6507
	$\hat{Y}^{SH}_{DESPREG}$	10,969,194	0.3630	6,525,294	0.2318	4,599,041	0.1733
$\hat{Y}e^{-2CV^2}$	$\hat{Y}^{SH}_{DE}$	9,876,572	0.3725	7,139,149	0.3148	5,482,075	0.2790
	$\hat{Y}^{SH}_{SP}$	11,045,714	0.4191	7,890,841	0.3650	5,992,191	0.3339
	$\hat{Y}^{SH}_{REG}$	10,753,319	1.8054	10,165,587	1.6800	9,838,704	1.6290
	$\hat{Y}^{SH}_{SPREG}$	14,346,129	0.4834	7,656,804	0.3221	5,747,293	0.2807
	$\hat{Y}^{SH}_{DESP}$	9,700,328	0.3403	6,996,714	0.2742	5,346,597	0.2320
	$\hat{Y}^{SH}_{DEREG}$	8,324,729	0.7021	7,746,781	0.6544	7,419,398	0.6344
	$\hat{Y}^{SH}_{DESPREG}$	10,627,667	0.3631	6,873,590	0.2702	5,202,075	0.2270

4.2.4. MSE와 MSPE를 이용한 비교

이 절에서는 MSE와 MSPE를 이용하여 각 소지역 추정량의 우수성을 비교하였다. 자료의 수에 따른 변화를 살펴보기 위해 표본 수는 3,000, 4,000 그리고 5,000개를 추출하여 각각 MSE와 MSPE를 비교하였다. 이에 관한 결과를 표 4.4에 작성하였다.

표 4.4를 살펴보면 먼저 자료수가 3,000인 경우 $\hat{Y}_{REG}$와 $\hat{Y}_{DEREG}$의 MSE가 가장 작은 것으로 나타나 이 두 추정량이 일반적으로 사용하고 있는 MSE를 기준으로 하고 표본 수가 작은 경우에는 우수한 결과를 주고 있음을 알 수 있다. 반면 MSPE를 기준으로 보면 위의 두 통계량이 가장 안 좋은 결과를 주고 있다. 이제 본 논문의 주된 내용인 축소 소지역 추정량을 살펴보자. 먼저 $\hat{Y}_{DE}$와 $\hat{Y}^{SH}_{DE}$는 두 축소 소지역 추정량 모두에서, MSE는 약 30% 감소를 MSPE는 약 50% 감소를 보이고 있다. 이런 결과는 다른 축소 소지역 추정량에서도 찾아볼 수 있다. 반면 $\hat{Y}_{REG}$와 $\hat{Y}_{DEREG}$의 MSE는 감소하지 않고 일정 수준을 유지하고 있으나 MSPE는 약 20%의 감소를 보이고 있다. 이는 축소 소지역 추정량의 우수성을 보여주는 것이다. 이러한 결과는 자료수가 증가했을 경우도 마찬가지이다. 먼저 자료수가 3,000에서 4,000으로 증가할 때, 모든 추정량에서 MSE와 MSPE가 감소하였다. 예를 들어 $\hat{Y}_{DE}$는 표본 수의 증가로 MSE와 MSPE가 약 40%정도 감소하였다. 이에 반하여 $\hat{Y}_{REG}$와 $\hat{Y}_{DEREG}$의 MSE와 MSPE는 상대적으로 작은 감소를 보이고 있다. 이제 축소 소지역 추정량을 살펴보자. 예를 들어 $\hat{Y}_{DE}$와 $\hat{Y}^{SH}_{DE}$를 비교해 보면 약 30%의 MSE가 감소하

였고 약 40%의 MSPE 감소를 보이고 있다. 이런 결과는 다른 추정량에서도 찾을 수 있다. 그러나 자료수가 3,000인 경우와 같이 $\hat{Y}_{REG}$와 $\hat{Y}_{DEREG}$의 MSE는 감소하지 않고 일정 수준을 유지하고 있는 반면 MSPE는 약 20%의 감소를 보이고 있다. 자료수가 5,000인 경우를 보면 자료수의 증가로 $\hat{Y}_{DE}$가 매우 우수한 결과를 주고 있다. 이에 비해 표본 수의 증가에도 불구하고 $\hat{Y}_{REG}$와 $\hat{Y}_{DEREG}$의 MSE는 상대적으로 작은 감소를 보이고 있다. 반면 MSPE의 경우 $\hat{Y}_{REG}$보다 $\hat{Y}_{DEREG}$가 더 많은 감소를 보이고 있다. 자료수가 5,000인 경우 $\hat{Y}\{(1+CV^2)/(1+3CV^2)\}$와 $\hat{Y}e^{-2CV^2}$의 결과에 약간의 차이를 보이고 있으나 결론적으로 $\hat{Y}\{(1+CV^2)/(1+3CV^2)\}$가 우수한 결과를 주고 있다. 이 통계량은 CV가 "0"에 가까운 경우 $\hat{Y}$로 매우 큰 CV에서는 $(1/3)\hat{Y}$가 되므로 이 구간에서 얻어진다. 반면 $\hat{Y}e^{-2CV^2}$인 경우 매우 큰 CV에서 "0"에 가까운 결과를 얻게 된다. 물론 본 논문에서 가정한 것은 CV가 작다는 것이므로 CV가 작은 경우에는 두 결과에는 차이가 없을 것이다. 이제 축소 추정량을 보면 전체적으로 MSE는 5–10%가 MSPE는 0–20%정도가 감소하였음을 확인 할 수 있다. 이상을 종합하여 보면 축소 소지역 추정량의 MSPE 뿐 아니라 MSE도 감소하였음을 알 수 있다.

5. 결론

소지역추정에 관한 많은 연구가 진행됨에 따라 여러 소지역 추정량이 제안되고 있으며 각 소지역 추정량에 대한 비교가 여러 면에서 검토되고 있다. 본 논문에서는 MSPE를 기준으로 한 축소 소지역 추정량을 제안하였으며 모의실험결과 MSPE가 큰 폭으로 감소하였음을 확인하였다. 또한 MSE도 감소함을 볼 수 있다. MSE의 감소는 축소 예측량에서 얻을 수 있는 일반적인 결과는 아니며 Park과 Shin (2006)의 결과를 살펴보면 큰 MSPE의 감소를 보이는 반면 MSE는 약간 증가하는 것을 알 수 있다. 즉 축소 소지역 추정량은 매우 중요한 기준인 MSE를 증가시키는 단점이 있으며, 모든 자료가 양수일 경우만 사용가능하다는 단점이 있다. 또한 축소 소지역 추정량은 불편 추정량이 아닌 것을 확인 할 수 있다. 이상의 장점과 단점에도 불구하고 MSPE를 사용할 수 있는 조건이 만족된다면, 축소 소지역 추정량은 높은 오차비율을 주는 작은 소지역의 추정에 사용될 경우 우수한 결과를 줄 수 있을 것으로 생각된다.

참고문헌

김달호, 김남희 (2002). 반복조사에서 소지역자료의 베이지안 분석, <응용통계연구>, **15**, 119–128.

신기일, 최봉호, 이상은 (2007). 공간통계 활용에 따른 소지역 추정법의 평가, <응용통계연구>, **20**, 229–244.

Fay, R. E. and Herriot, R. A.(1979). Estimates of income from small places: An application of James-Stein procedures to census data, *Journal of the American Statistical Association*, **74**, 269–277.

Kim, J. O. and Shin, K. I. (2006). Comparison of small area estimations by sample sizes, *The Korean Communications in Statistics*, **13**, 669–683.

Kim, Y. W. and Choi, H. A. (2004). Small area estimation techniques based on logistic model to estimate unemployment rate, *The Korean Communications in Statistics*, **11**, 583–595.

Park, H. and Shin, K. I. (2006). A shrinked forecast in stationary process favoring percentage error, *Journal of Time Series*, **27**, 129–139.

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

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

Shin, K. I. and Lee, S. E. (2003). Model-data based small area estimation, *The Korean Communications in Statistics*, **10**, 637–645.

[2007년 7월 접수, 2007년 10월 채택]

Shrinkage Prediction for Small Area Estimations*

Hee-Jin Hwang[1] Key-Il Shin[2]

Abstract

Many small area estimation methods have been suggested. Also for the comparison of the estimation methods, model diagnostic checking techniques have been studied. Almost all of the small area estimators were developed by minimizing MSE(Mean square error) and so the MSE is the well-known comparison criterion for superiority. In this paper we suggested a new small area estimator based on minimizing MSPE(Mean square percentage error) which is recently re-highlighted. Also we compared the new suggested estimator with the estimators explained in Shin *et al.* (2007) using MSE, MSPE and other diagnostic checking criteria.

Keywords: Small area estimation methods, spatial statistics, shrinkage predictor, MSPE(Mean square percentage error).

* This research was supported by the research fund of Hankuk University of Foreign Studies, 2007.

1) Graduate Student, Dept. of Statistics, Hankuk University of Foreign Studies, San89, Wangsan, Mohyun-dong, Yongin-si, Kyonggi-Do 449-791, Korea.
E-mail: lshhj01@naver.com

2) Corresponding author. Professor, Dept. of Statistics, Hankuk University of Foreign Studies, San89, Wangsan, Mohyun-dong, Yongin-si, Kyonggi-Do 449-791, Korea.
E-mail: keyshin@hufs.ac.kr