응용통계연구 (2009)
22(2), 403–414

MSPE를 이용한 임금총액 소지역 추정

황희진[1] · 신기일[2]

[1]한국외국어대학교 통계학과, [2]한국외국어대학교 통계학과

(2008년 12월 접수, 2009년 2월 채택)

요 약

국내외적으로 지역통계에 관한 관심이 높아지고 있으며 이와 관련하여 소지역 추정에 관한 많은 연구가 진행되고 있다. 소지역 추정에 사용되는 추정량의 대부분은 MSE(mean squared error)를 최소화하여 얻어진다 (Rao, 2003). 최근 황희진과 신기일 (2008)은 MSPE(mean squared percentage error)를 최소화하는 추정량을 사용한 소지역 추정법을 제안하였다. 본 논문에서는 노동통계 중 지청별 일인당 평균 임금총액 추정에 황희진과 신기일 (2008)이 제안한 방법을 적용하여 보았으며 2007년 매월 노동통계 자료를 이용하여 기존의 MSE를 최소화 하여 얻어진 여러 추정량과 우수성을 비교해 보았다. 또한 노동통계를 위한 소지역 추정의 실제 사용 가능성을 살펴보았다.

주요용어: 소지역통계, 공간통계, 축소추정량, 공간회귀추정량.

1. 서론

최근 지방자치제가 정착되어감에 따라 각 자치단체의 정책수립에 필수적인 지역통계 수요가 증가하고 있다. 하지만 대부분의 표본조사는 전국 규모로 설계되어 조사되고 있으며 별도의 지역별 통계를 작성하기 위한 노력은 아직 미미한 실정이다. 일반적으로 전국 규모 단위로 조사된 자료를 통해 원래 계획에 없던 소지역에 대해 직접추정을 하게 되면 그 소지역에 배정된 자료 수가 작기 때문에 필연적으로 분산이 매우 커져 원하는 수준의 정도를 얻기 어렵게 된다. 반면 모든 소지역 통계가 공표 수준의 정도를 유지하도록 표본설계를 하면 표본 규모가 커져 시간과 예산에 대한 부담이 매우 커지게 된다. 이러한 문제를 해결하기 위한 방법이 소지역 추정방법이다. 소지역 추정법은 여러 가지 제약조건 하에서 추정의 정도를 높이기 위한 다양한 방법을 제시하고 있는데 기존에 연구된 추정방법은 크게 자료기반 추정과 모형기반 추정으로 나누어진다. 일반적으로 설명력이 큰 보조변수가 존재한다면 모형기반 추정량이 자료기반 추정량보다 우수한 것으로 알려져 있다.

최근 국내에서도 소지역 추정에 관하여 많은 연구가 진행되고 있다. 특히 자료 분석에 초점을 맞춘 이계오와 이화영 (2008), 그리고 패널자료에 일반화추정방정식을 이용하여 소지역 추정을 연구한 여인권 등 (2008)은 소지역 추정의 필요성을 보여준다고 하겠다. 또한 모형기반 소지역 추정 방법을 연구한 논문으로는 김달호와 김남희 (2002) 그리고 Kim과 Choi (2004) 등이 있다. 본 논문에서는 모형기반 추정량을 중심으로 소지역 추정법을 살펴보았다. 특히 MSE를 최소화하여 얻어진 추정량 대신 최근 황희진과 신기일 (2008)이 제안한 MSPE를 최소로 하는 추정량을 살펴보았다. MSPE를 기준으로 한 추정량

이 논문은 2007년도 정부재원(교육인적자원부 학술연구조성사업비)으로 학술진흥재단의 지원을 받아 연구되었음(KRF-2007-313-C00124).

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

의 가장 큰 장점은 MSE를 기준으로 얻어진 기존의 추정량에 상수를 곱하여 쉽게 구한다는 것이다. 물론 MSPE를 사용하게 될 경우에는 고려해야 할 점이 많이 있다. 먼저 얻어진 자료가 모두 "0"보다 커야한다. 이는 모든 자료에 MSPE 기준을 적용할 수 없다는 점이 매우 큰 제약이라고 할 수 있다. 또한 "0"보다는 크더라도 "0"에 가까운 값을 갖는 자료의 경우에도 MSPE를 사용할 때는 매우 주의를 하여야 한다. 다음으로 자료의 범위(range)가 커야 한다. 자료의 범위가 작을 경우 MSPE를 사용하는 효과가 매우 떨어진다. 참고로 어떤 자료에 MSPE를 적용하면 효과가 있는지에 관한 설명은 Park과 Stefanski (1998)을 살펴보기 바란다. 따라서 이미 얻어진 많은 소지역 추정에 관한 결론들을 그대로 사용할 수 있게 된다. 또한 본 논문에서는 직접추정량의 변동성과 간접추정량의 편향을 동시에 줄이는 방법인 선형결합추정량도 고려하였으며 공간상관관계를 검토한 뒤 공간통계기법을 활용한 추정량도 살펴보았다. 결국 본 논문에서는 회귀추정량, 공간추정량, 공간회귀추정량 그리고 MSPE를 이용한 축소추정량과 이를 결합한 선형결합추정량 등 각 추정량의 우수성을 비교하였다. 또한 선형결합추정량을 만들 때 사용되는 가중치 α가 MSE 기준 또는 MSPE 기준을 사용하여도 변동이 없음을 증명하였다.

본 논문의 2절에서는 기존에 이미 보편적으로 사용되는 추정량과 MSPE 기준 추정량에 대하여 간략하게 살펴보았고, 3절에서는 추정량 비교를 위한 여러 가지 비교통계량을 소개하였다. 4절에서는 실제 자료를 이용하여 MSPE 기준 추정량과 기존의 추정량을 비교 분석하였다. 모든 자료 분석에서는 2007년 4월 특별 매월노동통계조사의 평균 임금총액 자료와 근로복지공단의 평균 임금총액 자료 그리고 고용정보원의 고용 DB 자료를 이용하였다.

2. 소지역 추정량

다양한 소지역 추정량이 제안되어 왔는데 본 논문에서는 MSE와 MSPE를 기준으로 하여 얻어진 여러 소지역 추정량을 실제 노동부 자료에 적용하였을 때 어떤 추정량이 우수한지 또는 적용 가능한지를 살펴보았다. 이에 이 절에서는 기존에 제안된 여러 추정량을 간단히 살펴보았다.

2.1. MSE를 이용한 소지역 추정량

먼저 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 추정하는 방법이다. 복합추정법은 직접추정량의 변동성을 작게 하고 합성추정량의 편향을 줄이기 위하여 두 추정량을 선형결합하여 얻는 추정방법이다. 이 방법들은 여인권 등 (2008), 이계오와 이화영 (2008), 황희진과 신기일 (2008) 등에서 자세히 살펴볼 수 있으며 본 논문에서는 이중에서 직접추정량만을 살펴보았다.

모형기반 추정법으로는 회귀분석방법, 경험적베이지안(empirical Bayesian: EB) 추정법, 계층적베이지안(hierarchical Bayesian: HB) 추정법 등이 있으며 일반적으로 계층적베이지안 추정법이 우수한 것으로 알려져 있다. 모형기반 추정량에 관한 자세한 내용은 김달호와 김남희 (2002)와 Kim과 Choi (2004) 등을 살펴보기 바라며, 본 논문에서는 이중에서 가장 많이 사용되고 있는 회귀분석 추정량을 사용하여 분석하였다. EB와 HB를 사용하기 위해서는 개별단위(unit level) 자료가 있어야 하는데 본 논문에서 사용된 자료는 지역단위(area level) 자료이므로 이 두 추정법의 사용이 불가능하였다. 반면에 공간상관관계가 있는 경우 이를 이용하여 분석하게 되면 더욱 우수한 결과를 얻을 수 있기 때문에 공간소지역 추정량도 본 논문에서 살펴보았다. 이제 본 논문에서 사용할 MSE를 최소로 하는 소지역 추정량을 정리하면 다음과 같다.

2.1.1. 직접추정량: $\hat{Y}_{DE}$

$$\hat{Y}_{DE} = \hat{Y}_i = \sum_j w_{ij} y_{ij}, \tag{2.1}$$

여기서 $\hat{Y}_i$는 i번째 소지역 추정값을 의미하고 w_{ij}는 추출 가중치를, y_{ij}는 i지역 j번째 자료를 나타낸다. 다만 본 논문에서 사용한 직접추정량은 총계 추정이 아닌 일인당 평균 임금총액을 추정하므로 일반적으로 사용하는 식 (2.1)과는 다른 식 (4.1)의 형태를 취하게 된다.

2.1.2. 회귀추정량: $\hat{Y}_{REG}$

$$\hat{Y}_{REG} = \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki}, \tag{2.2}$$

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

2.1.3. 공간추정량: $\hat{Y}_{SP}$

$$\hat{Y}_{SP} = \hat{Y}_i = \hat{\rho} S_i, \tag{2.3}$$

여기서 $\hat{Y}_i$는 i번째 소지역 추정값이며 S_i는 i번째 소지역의 이웃지역에서 얻어진 자료의 합 또는 평균이다. 본 논문에서는 평균을 사용하였다. 그리고 $\hat{\rho}$는 절편이 없는 회귀모형을 적합하여 얻은 추정된 계수이다.

2.1.4. 공간회귀추정량: $\hat{Y}_{SPREG}$

$$\hat{Y}_{SPREG} = \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki} + \hat{\rho} S_i, \tag{2.4}$$

여기서 설명변수 $x_{1i}, x_{2i}, \ldots, x_{ki}$는 식 (2.2)와 동일하며 S_i는 식 (2.3)에서 정의한 이웃자료의 평균이다. 이에 관한 내용은 황희진과 신기일 (2008)을 참고하기 바란다.

2.1.5. 선형결합추정량: $\hat{Y}_{DESPREG}$

$$\hat{Y}_{DESPREG} = \alpha_{SPREG} \hat{Y}_{DE} + (1 - \alpha_{SPREG}) \hat{Y}_{SPREG}, \tag{2.5}$$

여기서 가중치 α_{SPREG}는 $\hat{Y}_{DESPREG}$의 MSE를 최소로 하는 값을 사용한다. 즉 가중치 공식

$$\alpha_{SPREG} = \frac{\mathrm{MSE}(\hat{Y}_{SPREG})}{\mathrm{MSE}(\hat{Y}_{DE}) + \mathrm{MSE}(\hat{Y}_{SPREG}) + 2E(\hat{Y}_{DE} - Y)(\hat{Y}_{SPREG} - Y)} \tag{2.6}$$

를 사용한다. 그러나 $E(\hat{Y}_{DE} - Y)(\hat{Y}_{SPREG} - Y)$는 $\mathrm{MSE}(\hat{Y}_{SPREG})$ 또는 $\mathrm{MSE}(\hat{Y}_{DE})$에 비해 작아 일반적으로 무시될 수 있다. 따라서

$$\alpha_{SPREG} = \frac{\mathrm{MSE}(\hat{Y}_{SPREG})}{\mathrm{MSE}(\hat{Y}_{DE}) + \mathrm{MSE}(\hat{Y}_{SPREG})}$$

가 된다. 이제 $\hat{\alpha}_{SPREG}$를 구하기 위해서는 MSE를 추정하여야 한다. 그러나 많은 추정량이 불편성을 만족하거나 또는 MSE 추정이 쉽지 않기 때문에 MSE 대신 분산이 사용된다. 따라서 실제 자료 분석에는

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

가 사용된다. 본 논문에서는 편의상 가중치 $\alpha = 0.5$로 하여 선형결합추정량을 구하였다. 이에 관한 자세한 내용은 Rao (2003)을 살펴보기 바란다.

2.2. MSPE를 이용한 소지역 추정량

2.2.1. 축소예측량(shrinkage predictor: $\hat{Y}^{SH}$) 기존의 추정량을 구하는 기준은 MSE(mean squared error), 즉 $E(Y-\hat{Y})^2$을 최소화하는 것이었다. 이 기준은 Y값에 상관없이 같은 크기의 오차를 가정한 모형이 타당할 때 사용하는 일반적인 것이다. 그러나 상대적인 오차의 크기가 중요한 경우 큰 Y값에 비해 작은 Y값은 상대적으로 오차가 매우 클 수 있다. 이 경우 MSE를 기준으로 추정량을 구하고 또한 MSE로 추정량의 우수성을 파악하는 것은 문제가 될 수 있다. 따라서 이러한 경우 MSE 대신 MSPE(mean squared percentage error)를 기준으로 하여 추정량을 구하는 것이 대안이 될 수 있다. 이에 관한 내용은 Park과 Stefanski (1998)를 참조하기 바란다. MSPE를 이용한 예측량은

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

을 만족하는 $\hat{Y}$로 구해지며 Jeong과 Shin (2008)에서와 같이 평균과 분산에 대하여 다음을 가정하자.

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

이제 Taylor 전개를 사용하면

$$\hat{Y}_1^{SH} = \hat{Y}\left(1-2\mathrm{CV}^2\right) \tag{2.7}$$

또는

$$\hat{Y}_2^{SH} = \hat{Y}\frac{\left(1+\mathrm{CV}^2\right)}{\left(1+3\mathrm{CV}^2\right)} \tag{2.8}$$

또는

$$\hat{Y}_3^{SH} = \hat{Y}\exp\left(-2\mathrm{CV}^2\right) \tag{2.9}$$

을 얻게 된다. 본 논문에서는 황희진과 신기일 (2008)의 결과에서 식 (2.8)이 가장 우수한 결과를 주는 것으로 나타났기 때문에 $\hat{Y}_2^{SH} = \hat{Y}^{SH}$를 사용하였으며 이 추정량을 축소추정량이라 부르겠다. 자세한 내용은 황희진과 신기일 (2008)을 살펴보기 바란다.

2.2.2. 축소선형결합추정량에서의 가중치(α) 2.1.5절에서 설명하였듯이 기존의 선형결합추정량에서 사용한 가중치 α의 추정량은 선형결합추정량의 MSE를 최소로 하는 값을 사용한다. 그러나 축소선형결합추정량을 위한 가중치의 추정량은 MSPE를 최소로 하는 추정량을 사용하는 것이 타당하므로 이 절에서는 MSPE를 최소로 하는 가중치를 구하였다. 가중치를 구한 결과 MSPE를 최소로 하는 가중치 추정량과 MSE를 최소로 하는 가중치의 추정량이 근사적으로 같았으며 이에 관한 증명은 다음과 같다.

증명: 두 변수 Y_A, Y_B의 선형결합추정량을 $Y_C = \alpha Y_A + (1-\alpha)Y_B$라 하고 Y를 참값이라 하자. 그리고 MSE를 최소로 하는 가중치를 α_{MSE}, MSPE를 최소로 하는 가중치를 α_{MSPE}라 하자. 그러면

$$\alpha_{MSE} = \frac{\mathrm{MSE}(Y_B)}{\mathrm{MSE}(Y_A)+\mathrm{MSE}(Y_B)+2E(Y_A-Y)(Y_B-Y)}$$

가 되며 $2E(Y_A-Y)(Y_B-Y)$가 $\mathrm{MSE}(Y_A)$, $\mathrm{MSE}(Y_B)$에 비해 무시할 정도로 작으면

$$\alpha_{MSE} = \frac{\mathrm{MSE}(Y_B)}{\mathrm{MSE}(Y_A) + \mathrm{MSE}(Y_B)}$$

이 된다. 따라서

$$\hat{\alpha}_{MSE} = \frac{\widehat{\mathrm{MSE}}(Y_B)}{\widehat{\mathrm{MSE}}(Y_A) + \widehat{\mathrm{MSE}}(Y_B)} = \frac{\frac{1}{n_B}\sum_{i=1}^{n}\left(Y_i^B - Y\right)^2}{\frac{1}{n_A}\sum_{i=1}^{n}\left(Y_i^A - Y\right)^2 + \frac{1}{n_B}\sum_{i=1}^{n}\left(Y_i^B - Y\right)^2} \tag{2.10}$$

이 된다. 여기서 Y는 참값이다. 이제 분모와 분자를 각각 Y^2으로 나누자. 그러면

$$\hat{\alpha}_{MSE} = \frac{\frac{1}{n_B}\sum_{i=1}^{n}\frac{(Y_i^B - Y)^2}{Y^2}}{\frac{1}{n_A}\sum_{i=1}^{n}\frac{(Y_i^A - Y)^2}{Y^2} + \frac{1}{n_B}\sum_{i=1}^{n}\frac{(Y_i^B - Y)^2}{Y^2}} \tag{2.11}$$

$$= \frac{\frac{1}{n_B}\sum_{i=1}^{n}\left(\frac{Y_i^B - Y}{Y}\right)^2}{\frac{1}{n_A}\sum_{i=1}^{n}\left(\frac{Y_i^A - Y}{Y}\right)^2 + \frac{1}{n_B}\sum_{i=1}^{n}\left(\frac{Y_i^B - Y}{Y}\right)^2} \tag{2.12}$$

이고 따라서 식 (2.12)는

$$\frac{\widehat{\mathrm{MSPE}}(Y_B)}{\widehat{\mathrm{MSPE}}(Y_A) + \widehat{\mathrm{MSPE}}(Y_B)} \tag{2.13}$$

가 된다. 여기서 $n_A = n_B = n$이다. 따라서

$$\hat{\alpha}_{MSE} = \hat{\alpha}_{MSPE}$$

을 얻는다.

본 논문에서는 MSPE를 최소로 하여 얻어진 가중치 추정량을 사용하지 않고 MSE를 기준으로 얻어진 추정량을 고려하였으며 논문을 간단히 하기 위하여 2.1.5절에서와 같이 $\hat{\alpha} = 0.05$를 사용하였다. □

2.3. 비교된 소지역 추정량

2.1절에서 소개된 소지역 추정량은 다음과 같다.

$$\hat{Y}_{DE},\quad \hat{Y}_{SP},\quad \hat{Y}_{REG},\quad \hat{Y}_{SPREG}\quad \text{그리고}\quad \hat{Y}_{DESPREG}.$$

다음으로 이 추정량을 이용하여 각각에 해당하는 축소 소지역 추정량을 정의할 수 있으나 황희진과 신기일 (2008)의 결과를 이용하여 이 중에서 가장 우수한 결과를 주는 다음의 축소추정량을 사용하여 분석하였다.

$$\hat{Y}_{REG}^{SH} = \hat{Y}_{REG}\frac{(1+\mathrm{CV}^2)}{(1+3\mathrm{CV}^2)},$$

$$\hat{Y}_{SPREG}^{SH} = \hat{Y}_{SPREG}\frac{(1+\mathrm{CV}^2)}{(1+3\mathrm{CV}^2)},$$

$$\hat{Y}_{DESPREG}^{SH} = \hat{Y}_{DESPREG}\frac{(1+\mathrm{CV}^2)}{(1+3\mathrm{CV}^2)}.$$

3. 비교통계량

소지역 추정량을 평가하기 위한 여러 비교통계량이 제안되었으며 그 중 본 논문에서 사용된 통계량은 R^2과 기울기 그리고 Rao (2003)에서 사용되고 있는 비교통계량이다. 이에 관한 자세한 내용은 Rao (2003)를 살펴보기 바라며 이절에서는 이것들을 간단히 설명하였다.

3.1. 회귀모형을 이용한 방법(R^2과 기울기)

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

3.2. 비교통계량

본 논문에서 사용한 비교 통계량은 소지역 i의 참값을 Y_i 그리고 각 소지역 추정량을 $\hat{Y}_i$라 했을 때 Mean Squared Error(MSE), Mean Squared Percentage Error(MSPE), Mean Absolute Error(MAE), Absolute Relative Error(ARE), Relative Bias(RB), Relative Efficiency(EFF)이며 다음과 같이 정의된다. 여기서 MSPE는 황희진과 신기일 (2008)과 같은 이름을 사용하였으며 다른 비교통계량은 Rao (2003)의 이름을 사용하였다.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)^2, \qquad \mathrm{MSPE} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{Y}_i - Y_i}{Y_i}\right)^2,$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{Y}_i - Y_i\right|, \qquad \mathrm{ARE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{Y}_i - Y_i}{Y_i}\right|,$$

$$\mathrm{RB} = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{Y}_i - Y_i}{Y_i}, \qquad \mathrm{EFF} = \left(\frac{\mathrm{MSE}(Y)}{\mathrm{MSE}(\hat{Y})}\right)^{\frac{1}{2}}.$$

4. 자료 분석 및 추정량 비교

4.1. 자료 분석

본 논문에서 사용한 자료는 2007년 4월 매월노동통계조사에서 사용되는 표본 약 7,000개에 특별표본(지역표본)을 더한 약 10,000개의 자료이며 관심변수는 산업대분류별, 규모별, 지청별 평균 임금총액이다. 매월노동통계조사는 전국단위 조사로 표본설계당시에는 지청별 통계가 아닌 시도별 통계만을 고려하기 때문에 대분류별, 지청별, 규모별 직접추정량을 구하면 해당 분류의 자료 수 부족으로 분산이 매우 크게 된다. 따라서 이러한 직접추정의 단점을 보완하기 위하여 임금총액과 관련된 의미 있는 보조변수를 찾고 이용 가능성을 검토하였으며, 그 중 근로복지공단의 고용보험 자료가 적합하다고 판단되어 이를 이용한 회귀모형을 적합하였다. 또한 지청 자료에 대한 공간상관관계를 구해보고 공간모형 및 공간변수와 보조변수를 함께 고려한 공간회귀모형의 적용 가능성을 검토하였다. 그리고 이들 추정값에 대하여 3절에서 살펴본 비교통계량을 통해 효율성을 비교하였다.

본 논문에서는 대분류별, 사업체 규모별로 구성된 모든 셀을 분석할 수 없기 때문에 대표적인 두 가지

경우를 살펴보았다.
CASE1 : 제조업(C), 규모 1(상용종사자 수 5~9인)
CASE2 : 제조업(C), 규모 3(상용종사자 수 30~99인).

4.1.1. 직접추정량: $\hat{Y}_{DE}$

$$\hat{Y}_{DE} = \hat{Y}_i = \frac{\sum_j^{n_i} w_{ij} y_{ij}}{\sum_j^{n_i} w_{ij} e_{ij}} = \sum_j^{n_i} w_{ij}^* y_{ij}, \tag{4.1}$$

여기서 $\hat{Y}_i$는 i번째 소지역(지청)의 평균 임금총액 추정값을 의미하고 w_{ij}는 추출 가중치를, y_{ij}는 i지역 j번째 사업체 자료를, e_{ij}는 i지역 j번째 사업체의 종사자수를 나타낸다. 이 식은 식 (2.1)과 같은 형태이다.

4.1.2. 회귀추정량: $\hat{Y}_{REG}$

$$\hat{Y}_{REG} = \hat{Y}_i = \hat{\beta}_1 x_{1i}, \tag{4.2}$$

여기서 x_{1i}는 보조변수이고 $\hat{\beta}_1$은 추정된 회귀계수이다. 유용한 보조변수를 찾기 위해 평균 임금총액 변수와 관련 있는 여러 변수를 검토한 결과 근로복지공단의 고용보험 자료와 한국고용정보원의 사업장 고용 DB 자료를 연결하여 얻은 사업장별 임금총액자료를 보조변수로 선택하였다. 이 자료는 자료 수도 많고 우편번호를 통한 지역 정보, 사업체 규모, 산업분류 등의 필요한 정보가 모두 포함되어 있어 설명변수로 적합하다고 판단하였기 때문이다. 따라서 고용보험 자료의 평균 임금총액을 설명변수로, 그리고 식 (4.1)에서 구한 직접추정량을 종속변수로 하는 회귀모형을 적합하여 식 (4.2)와 같은 회귀추정량을 구하였다. 아쉬운 것은 노동부 자료와 고용보험 자료에 대해서, 사업장번호와 같이 사업체 단위로 연결할만한 정보가 있었더라면 지역 단위(area level) 모형뿐 아니라 개별 단위(unit level) 모형을 추가적으로 분석하여 좀 더 정도 있는 추정량을 얻을 수 있다는 점이다. 한편 여기서 절편 없는 회귀모형을 적합한 것은 종속변수와 독립변수는 사실상 같은 평균 임금총액 자료이므로 원점을 통과해야 한다고 판단했기 때문이다. 결과적으로 논문에서 사용된 자료값이 모두 큰 값을 갖고 있으므로 기울기와 R^2은 "1"에 매우 가깝게 추정되었다. 즉 기울기와 R^2이 "1"에 가깝게 추정되는 것은 두 자료값이 매우 일치하기 때문에 일어나는 현상이 아닐 수도 있으며 두 자료값이 모두 크고, 절편이 없는 회귀모형을 사용했기 때문일 수도 있는 것에 주의할 필요가 있다.

4.1.3. 공간추정량: $\hat{Y}_{SP}$

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

앞에서 언급한 고용보험 자료를 이용한 회귀모형은 설명력이 매우 높다. 다만 회귀분석 결과를 보면 Root MSE, MAE 등의 값이 큰 편이다. 이 때 만약 임금총액 변수가 서로 공간상관관계를 가지고 있다면 공간정보를 이용하는 것이 타당할 것이다. 이를 위하여 공간상관관계를 나타내는 Moran's I를 구하였으며 이 때 이웃정보는 경계를 공유하는 경우를 이웃으로 정의하여 사용하였다. 각 CASE 별로 구한 Moran's I는 다음과 같다.

표 4.1을 보면 두 CASES 모두 Moran's I 값이 유의하게 나타나 공간상관관계가 존재하는 것으로 판단할 수 있다. 따라서 평균 임금총액 자료에 대하여 공간상관관계가 존재하므로 이웃지역 자료의 평균을 보조변수로 하는 절편없는 회귀모형을 적합하여 공간추정량을 구하였다.

표 **4.1.** 공간상관관계(MORAN'S I)

CASES	Moran's I	P-값
CASES1	0.2935	0.0001
CASES2	0.4067	0.0001

표 **4.2.** 추정결과

추정량		Root MSE	MAE	$\hat{\beta}_1$	$\hat{\rho}$	R^2
CASE1	$\hat{Y}_{SP}$	252.860	280.707	-	0.98830(0.0001)	0.9674
	$\hat{Y}_{REG}$	224.806	176.190	1.13464 (0.0001)	-	0.9848
	$\hat{Y}_{SPREG}$	225.672	176.190	1.28617 (0.0001)	−0.13410(0.4150)	0.9850
CASE2	$\hat{Y}_{SP}$	379.599	280.707	-	0.93750(0.0001)	0.9654
	$\hat{Y}_{REG}$	238.134	184.271	1.08758 (0.0001)	-	0.9864
	$\hat{Y}_{SPREG}$	235.803	189.931	0.93677 (0.0001)	0.13997(0.1805)	0.9870

∗ ( )안은 추정값에 대한 p-value

4.1.4. 공간회귀추정량: $\hat{Y}_{SPREG}$

$$\hat{Y}_{SPREG} = \hat{Y}_i = \hat{\beta}_1 X_i + \hat{\rho}\bar{S}_i.$$

회귀모형과 공간모형이 각각 의미있게 나타나므로 두 변수를 모두 설명변수로 하여 공간회귀모형을 적합하였다. 앞에서와 마찬가지로 절편이 없는 모형을 사용하였다.

4.1.5. 축소추정량: $\hat{Y}^{SH}_{REG}, \hat{Y}^{SH}_{SPREG}, \hat{Y}^{SH}_{DESPREG}$ 축소추정량은 CV를 이용해서 간편하게 구할 수 있을 뿐 아니라 MSPE 기준에서 가장 높은 효율성을 보인다. 이제까지 구한 추정량과 각 추정량의 CV를 이용하여 각 추정량에 대한 축소추정량을 구하였다.

4.2. 추정량 비교

이 절에서는 3절에서 설명한 비교통계량을 이용하여 각 소지역 추정량을 비교하였다.

4.2.1. 추정결과 표 4.2는 공간추정량, 회귀추정량 그리고 공간회귀추정량의 비교를 위해 각 모형을 적합하여 얻어진 추정 결과이다. 이를 살펴보면 공간추정량은 두 경우 모두 MAE나 Root MSE 값이 회귀추정량에 비해 크고 R^2도 작다. 또한 공간회귀추정량을 보면 공간변수 $\bar{S}_i$가 유의하지 않아 회귀추정량과 크게 차이나지 않음을 알 수 있다.

4.2.2. 기울기와 R^2 직접추정량 $\hat{Y}_{DE}$를 종속변수로 하고 다른 추정량을 독립변수로 하여 얻어진 기울기와 R^2을 비교한 결과가 표 4.3에 나와 있다. 공간추정량, 회귀추정량 그리고 공간회귀추정량의 기울기는 이론적으로 "1"이 되어야 하며 이에 관한 증명은 신기일 등 (2007)에 나와 있다. 따라서 추정된 기울기가 "1"보다 많이 작으면 나쁜 추정량으로 판단할 수 있으나 기울기가 "1"이라고 우수한 추정량이라 할 수 없다. 여기서 축소추정량들은 원 추정량들보다 기울기가 커지게 되는데 이는 축소추정량이 불편추정량이 아니라 약간의 편향을 갖고 있기 때문에 나타나는 특성이다. 결과를 살펴보면 "1"에서는 크게 벗어나지 않았으며 R^2도 매우 높아 소지역 추정량으로 사용가능하다고 판단할 수 있다.

4.2.3. 추정량의 비교 3.2절에서 설명한 비교통계량을 이용하여 각 추정량의 효율성을 검토해보았다. 일반적으로 MSE를 최소로 하는 추정량 중에서는 EB 또는 HB가 가장 우수한 결과를 주는 것으로 알려

표 **4.3.** 추정량별 기울기와 R^2

추정량			기울기	R^2
CASE1	$\hat{Y}$	$\hat{Y}_{SP}$	1.00000	0.9674
		$\hat{Y}_{REG}$	1.00000	0.9848
		$\hat{Y}_{SPREG}$	1.00000	0.9850
		$\hat{Y}_{DESPREG}$	1.00379	0.9963
	$\hat{Y}^{SH}$	$\hat{Y}^{SH}_{REG}$	1.01959	0.9848
		$\hat{Y}^{SH}_{SPREG}$	1.02250	0.9850
		$\hat{Y}^{SH}_{DESPREG}$	1.03190	0.9963
CASE2	$\hat{Y}$	$\hat{Y}_{SP}$	1.00000	0.9654
		$\hat{Y}_{REG}$	1.00000	0.9864
		$\hat{Y}_{SPREG}$	1.00000	0.9870
		$\hat{Y}_{DESPREG}$	1.00329	0.9967
	$\hat{Y}^{SH}$	$\hat{Y}^{SH}_{REG}$	1.04568	0.9864
		$\hat{Y}^{SH}_{SPREG}$	1.03842	0.9870
		$\hat{Y}^{SH}_{DESPREG}$	1.05233	0.9967

표 **4.4.** 비교통계량을 이용한 결과

추정량			MSE	MAE	ARE(%)	MSPE(%)	RB(%)	EFF(%)
CASE1	$\hat{Y}$	$\hat{Y}_{DE}$	11109.3	89.2	5.32	0.41	5.04	137.17
		$\hat{Y}_{SP}$	32328.3	146.6	8.41	1.05	0.15	80.41
		$\hat{Y}_{REG}$	21995.3	117.0	6.89	0.75	2.36	97.49
		$\hat{Y}_{SPREG}$	25016.5	125.0	7.34	0.84	2.45	91.41
		$\hat{Y}_{DESPREG}$	11779.7	91.6	5.45	0.43	3.74	133.21
	$\hat{Y}^{SH}$	$\hat{Y}^{SH}_{REG}$	20175.6	105.8	6.18	0.67	0.40	101.79
		$\hat{Y}^{SH}_{SPREG}$	22854.2	112.2	6.54	0.75	0.19	95.64
		$\hat{Y}^{SH}_{DESPREG}$	8105.5	74.4	4.38	0.28	0.92	160.59
CASE2	$\hat{Y}$	$\hat{Y}_{DE}$	48464.3	193.3	10.70	1.50	9.91	87.07
		$\hat{Y}_{SP}$	39056.0	149.8	8.72	1.40	5.16	97.00
		$\hat{Y}_{REG}$	30931.6	151.1	8.37	0.94	8.12	108.99
		$\hat{Y}_{SPREG}$	28216.6	147.7	8.22	0.87	8.09	114.12
		$\hat{Y}_{DESPREG}$	34354.2	165.6	9.18	1.06	9.00	103.42
	$\hat{Y}^{SH}$	$\hat{Y}^{SH}_{REG}$	12615.9	86.2	4.72	0.37	3.39	170.66
		$\hat{Y}^{SH}_{SPREG}$	12104.5	91.2	5.06	0.37	4.09	174.23
		$\hat{Y}^{SH}_{DESPREG}$	12243.8	99.9	5.57	0.38	3.92	173.24

져 있으며 또한 회귀추정량 등도 MSPE를 기준으로 얻어진 축소 소지역 추정량보다는 MSE를 기준으로 하였을 때 우수한 결과를 주어야 한다. 물론 이러한 결과는 MSE 추정량 등 위에서 설명된 비교통계량을 구하기 위한 참값이 존재하여야 얻어질 수 있다. 그런데 지청별 평균 임금총액의 참값을 구하기 위한 모집단 자료가 존재하지 않으므로 비교를 위해서 본 논문에서는 지청별 참값 대신 노동부 공표수준인 시도별 평균 임금총액을 참값으로 가정하였다. 결국 본 논문에서 얻어진 MSE 등의 비교통계량에서 얻어진 결과의 사용에는 주의를 하여야 한다

이제 비교통계량은 다음과 같이 나누어 볼 수 있다. 먼저 오차의 크기 비교를 위한 비교통계량으로는 MSE와 MAE가 있으며, 상대적인 오차의 크기를 비교하기 위한 것으로는 ARE와 MSPE가 있다. 또한 편향(bias)에 관한 것으로는 RB가 있고 마지막으로 상대적인 효율성을 비교하기 위한 것으로 EFF가 있다.

표 4.4에 각 추정량에 대하여 비교통계량 값을 구한 결과가 나와 있다. 먼저 CASE1의 경우 MSE를 보면 기존의 추정량들 중에서는 직접추정량 $\hat{Y}_{DE}$의 값이 11109.3으로 가장 작으며 회귀추정량 $\hat{Y}_{REG}$나 공간추정량 $\hat{Y}_{SP}$, 공간회귀추정량 $\hat{Y}_{SPREG}$는 비교적 큰 값을 나타내고 있다. 또한 직접추정량의 MSE 값이 작기 때문에 직접추정량과의 선형결합추정량 $\hat{Y}_{DESPREG}$도 좋은 결과를 보이고 있다. 한편 이들 추정량에 대한 축소추정량에 대하여 구한 값을 보면 모두 값이 줄어든 것을 알 수 있는데 특히 $\hat{Y}^{SH}_{DESPREG}$가 8105.5로 가장 작다. 비슷한 개념의 MAE 역시 같은 결과를 보이고 있는데 이렇듯 직접추정량의 오차가 작을 경우에는 직접추정량을 포함하는 선형결합추정량의 효율성이 높아지게 된다. 물론 가장 우수한 결과를 주는 것은 축소추정량이다.

CASE2의 경우 MSE와 MAE를 보면 기존의 추정량들 가운데 가장 작은 MSE 값을 가지는 것은 공간회귀추정량 $\hat{Y}_{SPREG}$이다. 이 경우에는 직접추정량 $\hat{Y}_{DE}$이 가장 큰 MSE 값을 나타내고 있으며 회귀추정량 $\hat{Y}_{REG}$이 우수한 결과를 주고 있다. 한편 공간추정량 $\hat{Y}_{SP}$의 MSE는 비교적 큰 값을 나타내고 있다. 이러한 경우 공간추정량 단독으로 사용하기 보다는 보조변수와 같이 사용하는 것이 타당하다. 따라서 공간회귀추정량을 사용하는 것이 좋은 방법이며 결과를 살펴보면 공간회귀추정량 $\hat{Y}_{SPREG}$의 정도는 매우 좋아지는 것으로 나타났다. 그리고 선형결합추정량 $\hat{Y}_{DESPREG}$의 MSE 역시 회귀추정량 $\hat{Y}_{REG}$나 공간회귀추정량 $\hat{Y}_{SPREG}$에 비해서는 큰 값을 보인다. 또한 CASE1과 마찬가지로 이 경우에도 축소추정량들은 기존의 추정량에 비해서 MSE 값이 대폭 줄어드는 것을 볼 수 있는데 그 결과 MSE가 가장 작은 것은 $\hat{Y}^{SH}_{SPREG}$이고 MAE가 가장 작은 것은 $\hat{Y}^{SH}_{REG}$이다.

다음으로 상대적 오차의 크기인 ARE와 MSPE를 살펴보면 역시 MSE와 거의 유사한 결과를 나타내고 있다. CASE1의 경우에는 $\hat{Y}^{SH}_{DESPREG}$가 가장 작은 값을 가지며 CASE2의 경우는 $\hat{Y}^{SH}_{REG}$의 값이 가장 작다.

RB는 상대적 편향의 크기로 직접추정량에 대하여 가장 작게 나타나는 것이 일반적이다. 하지만 여기서는 CASE1의 경우 공간추정량 $\hat{Y}_{SP}$의 값이 0.15%로 가장 작고 직접추정량 $\hat{Y}_{DE}$의 값이 5.04%로 가장 크게 나타나는데 이는 본 논문에서 구한 직접추정량과 참값으로 사용한 노동부 자료 사이에 추정 방식, 가중치 등으로 인한 차이가 존재하기 때문일 것으로 파악된다. CASE2의 경우에서도 $\hat{Y}_{DE}$가 큰 값을 갖고 있다. 두 CASES에서 모두 축소 추정량의 RB 값이 작아 축소추정량의 편향은 크지 않다고 판단된다.

마지막으로 상대적 효율성 비교를 위해 EFF를 살펴보면 CASE1의 경우 $\hat{Y}^{SH}_{DESPREG}$이 160.59%로 가장 높았으며, CASE2의 경우 $\hat{Y}^{SH}_{SPREG}$이 174.23%로 가장 높았다. 이는 참값의 MSE에 대한 상대적 효율성을 알아보기 위한 것으로 값이 클수록 우수한 추정량이 된다.

결론적으로 비교통계량을 살펴본 결과 CASE1처럼 직접추정량의 MSE 등의 값이 작고 효율성이 좋을 경우에는 직접추정량을 포함하는 선형결합추정량이 보다 나은 결과를 주는 반면, 직접추정량의 효율성이 떨어질 때에는 공간회귀추정량만을 사용하는 것이 우수한 결과를 나타낸다. 또한 직접추정량의 정도에 따라 어느 것이 가장 우수한 추정량이 될 수 있는가는 달라지지만 어느 경우에도 기존의 추정량보다는 축소추정량을 사용하게 되면 모든 통계량에서 값이 월등히 좋아지는 것을 알 수 있다. 특히 CASE1의 경우 $\hat{Y}^{SH}_{DESPREG}$의 MSE나 MSPE는 월등히 작아지고 있으며 CASE2의 경우도 축소추정량은 모두 높은 효율성을 보여주고 있다.

5. 결론

본 논문에서는 선형결합추정량을 구할 때 사용되어지는 가중치에 대하여 MSE를 기준으로 구한 값과 MSPE를 기준으로 구한 값이 결국 같아지는 것을 보였다. 또한 지청별 평균 임금총액에 대한 회귀추

정량, 공간추정량, 공간회귀추정량 그리고 직접추정량과의 선형결합추정량을 구해보았으며 각 추정량에 대한 축소추정량도 구하여 MSE, MSPE 등 비교통계량을 가지고 효율성을 비교하였다. MSE 기준 추정량들을 비교한 결과 직접추정량의 효율성이 좋을 경우에는 직접추정량과 직접추정량을 포함하는 선형결합추정량이 우수한 결과를 주는 반면, 직접추정량의 효율성이 떨어질 때에는 회귀추정량이나 공간회귀추정량만을 사용하는 것이 보다 나은 결과를 주었다. 그러나 어느 경우에도 MSPE를 기준으로 구해진 축소추정량은 모든 비교통계량에서 값이 월등히 좋은 것을 확인할 수 있었다. 이는 본 논문에서 사용된 자료인 평균 임금총액의 경우 각 지청별로 자료값에 차이가 커서 축소추정량의 사용에 적합한 자료이기 때문인 것으로 파악된다. 따라서 소지역 추정에서 축소추정량의 사용을 검토할 필요가 있다.

결론적으로 어느 추정량을 사용하는 것이 가장 바람직한지는 자료의 형태 및 조건에 따라 다를 수 있으며 이에 대한 상세한 검토와 판단이 필수적이다. 또한 EB와 HB 등의 우수한 소지역 추정법 사용을 위하여 개별 단위(unit level) 자료 확보가 반드시 필요하다.

참고문헌

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

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

여인권, 손경진, 김영원 (2008). 일반화추정방정식을 활용한 소지역 추정과 실업률 패널분석, <응용통계연구>, **21**, 665–674.

이계오, 이화영 (2008). 임금구조기본통계조사의 직업소분류별 임금추정에서 소지역추정법 적용방안 연구, <통계연구>, **13**, 237–256.

황희진, 신기일 (2008). 축소예측을 이용한 소지역 추정, <응용통계연구>, **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.

Kim, Y. W. and Choi, H. A. (2004). Small area estimation technique based on logistic model to estimate unemployment rate, *Communications of the Korean Statistical Society*, **11**, 583–595.

Park, H. and Stefanski, L. A. (1998). Relative-error prediction, *Statistics & Probability Letters*, **40**, 227–236.

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

A Small Area Estimation for Monthly Wage Using Mean Squared Percentage Error

Hee-Jin Hwang[1] · Key-Il Shin[2]

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

(Received December 2008; accepted February 2009)

Abstract

Many researches have been devoted to the small area estimation related with the area level statistics. Almost all of the small area estimation methods are derived based on minimization of mean squared error(MSE). Recently Hwang and Shin (2008) suggested an alternative small area estimation method by minimizing mean squared percentage error. In this paper we apply this small area estimation method to the labor statistics, especially monthly wages by a branch area of labor department. The Monthly Labor Survey data (2007) is used for analysis and comparison of these methods.

Keywords: Small area statistics, spatial statistics, shrinkage estimator, spatial regression estimator.

This research was supported by the Korea Research Foundation Grant funded by the Korean Government(MOEHRD, Basic Research Promotion Fund)(KRF-2007-313-C00124).

[2]Corresponding author: Professor, Department of Statistics, Hankuk University of Foreign Studies, San 89, Wangsan, Mohyun, Yongin, Kyonggi Do 449-791, Korea. E-mail: keyshin@hufs.ac.kr