연구논문

소지역 추정방법을 이용한 실업자 수 추정 사례연구*

Estimation of the Number of the Unemployed Using Small Area Estimation Methods

권세혁**
Kwon, Sehyug

정보화 사회에서는 목표지향적이고 세분화된 통계의 필요성이 높아지고 있으나 현재 사용되는 조사체계를 이용하면 추정 분산이 커져 생산된 통계의 정확도가 낮아진다. 표본크기를 늘리면 추정분산을 줄일 수 있으나 비용이나 시간 면에서는 비효율적이다. 현재와 비슷한 규모의 표본조사구 조사와 일반 행정통계를 이용하여 일정 신뢰수준을 갖춘 통계를 생산할 수 있는 소지역 추정법에 대한 연구가 진행되어 개발·적용되고 있다. 본 연구에서는 소지역 추정법을 활용하여 대전광역시의 5개 구별 실업자 수를 추정하고 추정치의 CV 값을 계산하여 추정방법의 효율성을 비교하는 사례분석을 실시하였다. 또한 합성추정량과 복합추정량의 MSE를 보다 정확하게 계산하는 방법으로 잭나이프 방법을 제안하고 계산방법을 보였다.

주제어 : 실업자 수, 소지역 추정법, 합성추정량, 복합추정량, 잭나이프

With the current sampling scheme, the sampling variance is getting larger in producing smaller regional statistics than the designed area. The larger sample size can make the variance reduced but the efficiency of sample survey lower. The desired confidence level of sampling survey can be obtained using the current sample scheme with the same sample size and administrative data. In this paper, the number of the unemployed of 5 regions in Daejon are estimated using small area estimation methods and the CV values in each estimation method is calculated and compared for their estimation efficiency as empirical study. Jackknife method is proposed to estimate the MSE of synthetic estimator and composite estimator more accurately.

Key words : number of the unemployed, small area estimation, synthetic estimation, composite estimator, Jackknife

* 이 논문은 2008년도 한남대학교 학술연구비에 의하여 연구되었음.

** 한남대학교 정보통계학과 교수 권세혁.
E - mail : wolfpack@hnu.ac.kr

Ⅰ. 서론

국가통계와 같이 국가 전체 구성원이 조사 모집단인 경우, 모집단 전체통계 또는, 광역 단위의 통계 생산을 목표로 하기 때문에 보다 좁은 영역이나 세분화된 단위의 통계 생산에는 어려움이 있다. 현재 통계청에서 매월 실시하고 있는 경제활동인구조사는 표본조사를 계획할 당시에는 전국 단위와 (광역)시·도 단위의 고용관련 통계 생산이 주목적이었으나, 1995년 지방자치제가 실시되면서부터 각 시·도 내의 시·군·구 단위의 통계 생산의 필요성이 꾸준히 제기되어 왔다. 그러나 현재 사용하는 경제활동인구조사 체계를 이용하여 시·군·구 단위의 고용관련 통계를 생산한다면, 각 시·군·구별로 배정된 조사구의 분포가 불균형적일 뿐만 아니라 어떤 시·군·구에는 한 개 또는, 두 개의 조사구가 배정되어 조사되기 때문에 각각의 고용통계에 대한 추정치의 분산이 커져 모수 추정치의 정확도는 매우 낮아지게 된다.

정보화 사회에서 요구하고 있는 통계는 지역적으로 좁은 영역을 의미할 뿐만 아니라 세분화된 범주 단위에 대한 고용통계도 요구되고 있다. (광역)시·도에서는 직업훈련이나 전문 인력 양성계획을 수립하는 데 있어서 직업별 고용통계가 필요하나 직업별로 배정된 표본조사구의 분포가 불균형적이기 때문에 현 조사체계로는 동일수준의 정확도를 갖는 직업별 실업률이나 경제활동참가율 등을 생산할 수 없다. 시·군·구 단위의 고용통계를 생산하기 위하여 표본조사구를 증가시키면 조사원과 코딩 및 분석 관련 연구원의 증가와 통계 생산에 필요한 소요 기간이 길어지므로 효율적이지 못하다.

현재와 비슷한 규모의 표본조사구를 조사하더라도 인구주택총조사의 자료 또는, 주민등록인구 자료와 같은 일반 행정통계를 이용하여 어느 정도의 신뢰수준을 갖춘 통계를 생산할 수 있는 소지역 추정법(small area estimation)이 미국이나 캐나다 등에서 개발되고 있고 일부 분야에서는 적용되고 있다. 현재 많이 적용되고 있는 복합추정법은 안정적이고 추정오차도 작다는 장점을 갖는 대신, 그것이 비편향 추정법이고 추정오차를 나타내는 평균제곱오차를 산정하는 적합한 방법이 없다는 것이 단점이다. 그러므로 이에 대한 연구가 선행되어야 시·군·구의 고용통계를 생산하는 데 소지역 추정법을 적용할 수 있을 것이다. 소지역 추정에 대한 방법론은 통계청(2000) 연구보고서에 정리되어 있으며, 고용통계에 관한 소지역 추정법에 대한 연구로는 충북 시·군·구 실업자 추정 연구(이계오 2000; 정연수 외 2003)와 실업률 추정 중심으로 고용통계 생산을 위한 소지역 추정법 연구(김수택 외 2008) 등이 있다.

본 연구는 소지역 추정법을 이용하여 실업자 수를 추정하는 사례분석으로 다음과 같이 구성하였다. 2절에는 소지역 추정방법을 소개하고 합성추정량과 복합추정량의 MSE를 보다 정확하게 얻는 잭나이프 방법을 설명하였다. 3절에는 대전광역시의 5개 구별 실업자 추정(2007년 6월 발표자료)을 사례로 하여 소지역 추정치를 얻고 각 추정방법의 표준오차를 구하는 과정과 계산결과를 정리하였고, 추정방법의 효율성 비교를 위해 CV 값을 계산하여 복합추정방법이 가장 효율적인 방법임을 보였다. 그리고 2절에서 설명된 잭나이프 방법에 의해 합성추정량과 복합추정량의 MSE를 추정하여 정리하였다. 4절에는 본 연구의 결론과 향후연구에 대해 요약하였다.

II. 소지역 추정방법

소지역 추정법에는 설계기반 추정법(design-based estimation), 간접 추정법(indirect estimation) 그리고, 모형기반 추정법(model-based estimation) 등이 있다. 소지역 통계 작성 시 설계기반 추정량이 목표 요구정도를 만족한다면 우선적으로 설계기반 추정량을 이용하게 되며, 그렇지 못할 경우에는 추정량의 정확도를 확보할 수 있는 다른 추정법을 찾아야 한다.

설계기반 추정법은 직접추정량(direct estimator)과 수정된 직접추정량(modified direct estimator)으로 구분된다. 관심변수와 밀접한 관련이 있는 보조정보가 있는 경우에 이를 이용하는 사후층화추정량(post stratified estimator), 비추정량(ratio estimator), 회귀추정량(regression estimator) 등도 직접추정량이다. 직접추정량은 편향이 없는 추정량이지만 해당 소지역에 배정된 표본의 크기가 작은 경우에는 추정량의 분산이 커져서 신뢰성이 떨어지게 된다. 한편 수정된 직접추정량은 해당 소지역 이외의 다른 지역의 조사결과를 추정과정에 추가적으로 이용하며 추정량의 불편성은 근사적으로 유지된다.

직접추정량은 해당 소지역에서 조사된 자료만을 이용하여 추정되며 간혹 센서스나 행정자료로부터 획득된 보조정보를 추가하여 추정하기도 한다. 총계추정에 대한 직접추정량으로서 가장 간단한 것은 다음의 단순추정량(expansion estimator)이다.

$$\hat{Y}_{e,a} = \sum_{i \in s_a} \omega_i y_i \tag{1}$$

여기에서 s_a 는 소지역 a 의 표본들의 집합, ω_i 는 조사단위 i 에 대한 가중치를 나타낸다.

식(1)의 직접추정량은 불편추정량이나 소지역 a 의 표본크기가 작을 경우에는 분산이 커지기 때문에 신뢰성에 문제가 있을 수 있다.

간접추정법에는 합성추정량(synthetic estimator), 복합추정량(composite estimator), 표본크기 의존 복합추정량(sample size dependent estimator) 등이 있다. 해당 지역의 조사 그리고, 자료뿐만 아니라 해당 지역을 포함하고 있는 더 큰 지역의 조사자료를 소지역 추정과정에 이용하여 소지역 추정의 신뢰성을 확보하는 방법이다.

합성추정법(synthetic estimation)은 소지역 추정 시 소지역을 포함하는 대영역의 정보를 함께 이용하는 방법으로서 소지역과 대영역의 특성 구조가 유사하다는 가정 하에서 이용된다. 합성추정량의 분산은 직접추정량의 분산에 비해 작으나 전제한 가정이 성립하지 않을 경우에는 심각한 편향이 발생할 수 있다. 소지역의 특성치 평균이 전체 지역의 특성치 평균과 같다는 가정 하에서 만들어진 가장 간단한 형태의 합성추정량은 다음과 같다.

$$\hat{Y}_{sym,m,a} = N_a \frac{\sum_{i \in s} \omega_i y_i}{\sum_{i \in s} \omega_i} = N_a \bar{y} \tag{2}$$

복합추정량은 직접추정량의 불안정성과 합성추정량의 잠재적 편향 가능성을 보완하기 위해 두 추정량의 가중평균을 취하며 일반적인 형태는 다음과 같이 주어진다.

$$\hat{Y}_{com,a} = \lambda_a \hat{Y}_{dir,a} + (1-\lambda_a)\hat{Y}_{sym,a} \tag{3}$$

가중치 λ_a 는 0과 1 사이의 값으로 결정하는 방법은 크게 세 가지 정도로 구분될 수 있다. 첫 번째 방법은 가장 간단한 방법으로서 가중치 λ_a 을 고정계수로 두는 방법인데 추정량의 신뢰성에 문제가 있어 많이 사용되지는 않는다. 두 번째 방법은 추정하고자 하는 소지역의 표본크기를 반영하는 방법이다. 이 경우 가중치 λ_a 는 $\hat{N}_{e,a}/N_a$ 의 함수로 표현된다. Drew et al.(1982)은 표본크기에 의존하는 복합추정량으로서 다음과 같은 추정량을 제안했다.

$$\hat{Y}_{ssd,r,a} = \lambda_a \hat{Y}_{r,a} + (1-\lambda_a)\hat{Y}_{sym,r,a} \tag{4}$$

여기에서 $\lambda_a = \begin{cases} 1 & , \text{ if } \hat{N}_{e,a} \geqq \delta N_a \\ \hat{N}_{e,a}/\delta N_a & , \ otherwise \end{cases}$ 이며 N_a는 소지역 a의 모집단 크기이다. 소지역 a의 표본크기를 n_a라 하고 $n = \sum_i n_i$일 경우, $\hat{N}_{e,a} = N\frac{n_a}{n}$이고 δ 는 합성추정량 부분의 편향을 보정하기 위해 주관적으로 결정되는 값이다. 캐나다 노동력조사에서는 $\delta = 2/3$ 를

이용한다(Statistics Canada 1998). 식(4)의 복합추정량은 직접추정량의 정확도가 완전히 확보된 지역에 대해서는 합성추정량의 가중치가 0 이 되기 때문에, 이러한 지역의 경우 직접추정값이 곧 복합추정값으로 선택된다고 볼 수 있다. 그렇지 않은 기타 지역에 대해서는 직접추정값과 합성추정값의 가중평균값으로 복합추정값이 계산된다. 캐나다 노동력조사에서 이러한 기타 지역들에 대한 합성추정량의 평균가중치는 약 10% 정도이며 많아도 20% 를 초과하지는 않는다. 이때 δ 의 값은[2/3 , 3/2] 의 범위에 있는 것으로 알려져 있다.

Sandal(1984)는 표본크기 의존 복합추정량과 가중치를 제안하였으며, Rao(2003)는 다른 형태의 가중치를 제안하고 정리하였고 소지역 추정량의 MSE 추정방법도 기술하였다. 소지역 추정법에서 잭나이프법에 의한 MSE추정에 대한 연구는 Jiang & Lahiri(2002)에 의해 제안되었다. 잭나이프 반복 값은 한 번에 하나의 소지역을 빼놓고 계산과정에 따라 추정값을 산정하게 되어 m개 소지역에 대응하는 m개의 반복 값이 얻어진다. 전체표본과 m개 반복표본으로부터 모형의 모수들(σ_b^2과 β)을 최우추정법, 적률법 또는 상수적합법 등 적절한 추정법을 이용하여 추정한 후, 이들로부터 평균제곱오차의 추정값을 산정한다. 표본오차인 σ_{ei}^2은 사전에 알고 있는 값으로 가정한다.

절차1. 모든 표본을 이용하여 $\hat{\sigma_b^2}$과 $\hat{\beta}$을 추정하고, k 번째 소지역의 자료를 제외하고 나머지 $(m-1)$개 표본을 이용하여 $\widehat{\sigma_{b(k)}^2}$과 $\widehat{\beta_{(k)}}$를 산정한다.

절차2. $\hat{\gamma_i} = z_i^2\hat{\sigma}_b^2(z_i\hat{\sigma}_b^2+\sigma_{ei}^2)^{-1}$,

$\hat{\gamma}_{ik} = z_i^2\hat{\sigma}_{b(k)}^2(z_i^2\hat{\sigma}_{b(k)}^2+\sigma_{ei}^2)^{-1}$,

$\hat{Y}_i^H = \hat{\gamma}_i\hat{Y}_i^D+(1-\hat{\gamma}_i)X_i^T\hat{\beta}$,

$\hat{Y}_{i(k)}^H = \hat{\gamma}_{i(k)}\hat{Y}_i^D+(1-\hat{\gamma}_{i(k)})X_i^T\hat{\beta}_{(k)}$를 계산한다.

절차3. $m_{1i(J)} = g_{1i}(\hat{\sigma}_b^2)-(m-1)\sum_{i=1}^{m}g_{1i}(\hat{\sigma}_{b(k)}^2)-g_{1i}(\hat{\sigma}_b^2)/m$을 계산한다.

절차4. $m_{2i(J)} = (m-1)\sum_{i=1}^{m}(\hat{Y}_{i(k)}^H-\hat{Y}_i^H)^2/m$을 계산한다.

절차5. 잭나이프 방법에 의한 EBLUP의 평균제곱오차에 대한 추정값은 $mse_{Jack}(\hat{Y}_i^H) = m_{1i(J)}+m_{2i(J)}$에서 산정한다.

위와 같은 절차를 통해서 계산한 $mse_J(\hat{Y}_i^H)$는 상당히 안정된 값을 갖는 특성을 갖고

있으나 계산과정이 복잡하다는 단점이 있다. SAS 버전 9.2에서는 소지역 잭나이프 추정이 가능한 프로시저가 제공되고 있다.

III. 실증분석

소지역 고용통계 생산을 위해 가장 널리 이용되는 방법은 현재 경제활동인구조사의 자료를 이용한 복합추정량이다. 소지역추정법으로 실업자 수를 추정하는 데 있어서도 타당한지를 검증하기 위해, 2007년 6월 대전광역시의 구별 실업통계를 작성하는 사례를 이용하여 알아보고자 한다. 대전광역시는 5개의 구를 가지고 있고, 매월 경제활동인구조사를 위해서 81개 조사구를 조사하고 있다. 각 구별 조사구 수와 경제활동인구와 실업자 수를 <표 1>에 정리하였다.

<표 1> 대전광역시 구별 경제활동인구 (2007년 6월)

구분 (조사구 수)		항목 (단위: 명)			
		경제활동인구	취업자	실업자	비경제활동인구
동구 (15)	남자	259	251	8	74
	여자	178	173	5	194
	계	437	424	13	268
중구 (16)	남자	261	257	4	88
	여자	167	165	2	229
	계	428	422	6	317
서구 (27)	남자	350	330	20	175
	여자	297	291	6	302
	계	647	621	26	477
유성구 (9)	남자	112	110	2	64
	여자	85	82	3	103
	계	197	192	5	167
대덕구 (14)	남자	183	173	10	71
	여자	120	118	2	126
	계	303	291	12	197
합계 (81)	남자	1165	1121	44	472
	여자	847	829	18	954
	계	2012	1950	62	1426

1. 직접추정량

$$\widehat{Y_i.} = \sum_{s=1}^{2} {}_s\widehat{Y_i.} \quad , \; i = 1,2,\cdots,I \; ; s = 1,2 \; ; h = 1,2,\cdots,n_i$$

$$= \sum_{s=1}^{2}\sum_{h=1}^{n_i} {}_s\widehat{Y_{ih}} = \sum_{s=1}^{2}\sum_{h=1}^{n_i} {}_sM_i \; {}_sY_{ih} \tag{5}$$

여기에서 s 는 성별을 나타내는 첨자, n_i 는 경제활동인구조사에서 i 번째 지역의 표본조사구 수, ${}_sY_{ih}$ 는 각 성별에 대해서 i 번째 지역의 표본조사구에서 조사한 실업자 수를 나타낸다. 승수 ${}_sM_i = {}_s\widehat{X_i.} / {}_sX_i.$ 은 $\widehat{Y_i.}$ 이 불편추정량이 되도록 산정한다. 여기에서 ${}_s\widehat{X_i.}$ 은 i 번째 지역에 대한 15세 이상의 상주 추계인구를 나타내며, ${}_sX_i.$ 는 경제활동인구조사에서 조사된 15세 이상의 조사인구를 나타낸다.

직접추정량 $\widehat{Y_i.}$ 의 분산은 다음 (6)식과 같고 분산의 추정량은 식(7)에 의해 계산된다.

$$Var(\widehat{Y_i.}) = \sum_{s=1}^{2} Var({}_s\widehat{Y_i.}) + 2\,Cov({}_1\widehat{Y_i.}\,,{}_2\widehat{Y_i.}) \quad , \; i = 1,2,\cdots,I$$

$$= \sum_{s=1}^{2} {}_sM_i^2\, Var(\sum_{h=1}^{n_i} {}_sY_{ih}) + 2\,{}_1M_i\,{}_2M_i\, Cov(\sum_{h=1}^{n_i} {}_1Y_{ih}\,,\sum_{h=1}^{n_i} {}_2Y_{ih}) \tag{6}$$

$$\hat{V}ar(\widehat{Y_i.}) = \sum_{s=1}^{2} {}_sM_i^2(\zeta_i \sum_{h=1}^{n_i} {}_sU_{ih}^2) + 2\,{}_1M_i\,{}_2M_i(\zeta_i \sum_{h=1}^{n_i} {}_1U_{ih}\,{}_2U_{ih}) \tag{7}$$

여기에서 ${}_sU_{ih} = d\,{}_sY_{ih} - {}_s\rho_i \cdot d\,{}_sX_{ih}$, $d\,{}_sY_{ih} = {}_sY_{ih} - {}_sY_{i,h+1}$, $d\,{}_sX_{ih} = {}_sX_{ih} - {}_sX_{i,h+1}$, ${}_s\rho_i = {}_sY_i. / {}_sX_i.$, $\zeta_i = [1 - n_i/(10N_i)]\, n_i/[2(n_i-1)]$ 이고, N_i 는 소지역 i 에 대한 모집단의 조사구 수를 나타낸다.

〈표 2〉는 취업자 수, 실업자 수, 경제활동인구에 대한 직접추정치와 추정치의 표준오차를 정리한 것이다. 대영역 표본설계에 기반을 둔 직접추정량은 각 소지역에 할당된 표본조사구의 수가 충분하지 않기 때문에 소지역 실업통계의 정확도를 제공하지는 못하므로 합성추정량이나 복합추정량에 비해 활용도가 적다.

〈표 2〉 대전시 소지역 직접추정량과 추정오차

(단위: 명)

구	취업자 수		실업자 수		경제활동인구	
	추정량	표준오차	추정량	표준오차	추정량	표준오차
동구	178,637	5,555	5,746	2,704	110,719	5,930
중구	174,781	8,214	2,460	828	130,635	7,639
서구	272,133	8,255	11,009	2,454	204,982	7,868
유성구	74,401	3,323	1,962	626	62,962	3,367
대덕구	125,412	4,928	5,210	1,703	85,192	5,201
총합계	825,364		26,387		594,490	

2. 합성추정량

대영역을 I개의 시·군·구 단위의 소지역들로 분할하고 대영역을 특성 기준에 따라 유사성을 갖는 J개의 성별 - 연령대별 범주들로 구분할 때, i 번째 소지역의 합성추정량 $\hat{Y}_{i.}^{S}$는 다음과 같이 주어질 수 있다.

$$\hat{Y}_{i.}^{S} = \sum_{j=1}^{J} \eta_{ij} \psi_{.j}^{a} \ , \quad i = 1,2,\cdots,I \ , \tag{8}$$

식(8)에서 가중치 $\eta_{ij} = (\xi_{ijt}^{C} / \xi_{ijt}^{R}) \, \xi_{ij} \, {}_{s}\kappa_{j}$ 는 i 번째 소지역에서 j 번째 범주에 대한 경제활동 추정인구를 나타낸다. 여기에서 ξ_{ijt}^{C}는 t 번째 해의 상주추정인구, ξ_{ijt}^{R}는 같은 해의 주민등록인구, ${}_{s}\kappa_{j}$는 경제활동인구조사에서 각 성별에 대한 j 번째 범주의 경제활동 참가율을 나타낸다. $\psi_{.j}^{a} = \hat{Y}_{.j.} / \sum_{i=1}^{I} \psi_{ij}$ $(j = 1,2,\cdots,J)$는 경제활동인구조사에서 추정된 j 번째 범주에 대한 실업률을 나타낸다. 본 연구에서는 경제활동인구에서 실업의 유사성을 고려하여 4개 범주 (남자, 15세~29세), (여자, 15세~29세), (남자, 30세 이상),(여자, 30세 이상)을 활용하여 추정량을 계산하였다.

소지역 i 에서 j 번째 범주에 대한 경제활동인구 η_{ij}를 상수로 가정한다면 합성추정량 $\hat{Y}_{i.}^{S}$의 분산은 다음 식(9)와 같다.

$$Var(\hat{Y}_{i.}^{S}) = \sum_{j=1}^{J} \eta_{ij}^{2} Var(\psi_{.j.}^{a}) + 2 \sum_{j<l} \eta_{ij} \eta_{il} Cov(\psi_{.j.}^{a}, \psi_{.l.}^{a}), \quad i = 1,2,\cdots,I \tag{9}$$

합성추정량 $\hat{Y}_{i.}^{S}$ 의 추정분산은 다음 식에 의해 계산될 수 있다.

$$\hat{Var}(\hat{Y}_{i.}^{S}) = \sum_{j=1}^{J} \eta_{ij}^{2} \left(\frac{1}{\sum_{i=1}^{I} \psi_{ij}} \right)^{2} \hat{Var}(\hat{Y}_{.j.}) + 2 \sum_{j<l} \eta_{ij} \eta_{il} \left(\frac{1}{\sum_{i=1}^{I} \psi_{ij}} \right) \left(\frac{1}{\sum_{i=1}^{I} \psi_{il}} \right) \hat{Cov}(\hat{Y}_{.j.}, \hat{Y}_{.l.})$$

$$= \sum_{j=1}^{J} \eta_{ij}^{2} \left(\frac{1}{\sum_{i=1}^{I} \psi_{ij}} \right)^{2} \left(M_j^{2} \zeta_j \sum_{i=1}^{I} \sum_{h=1}^{n_{ij}} U_{ijh}^{2} \right) + 2 \sum_{j<l} \eta_{ij} \eta_{il} \left(\frac{1}{\sum_{i=1}^{I} \psi_{ij}} \right) \left(\frac{1}{\sum_{i=1}^{I} \psi_{il}} \right) \left(M_j M_l \zeta_j \sum_{i=1}^{I} \sum_{h=1}^{n_{ij}} U_{ijh} U_{ilh} \right) \quad (10)$$

여기에서 $U_{ijh} = d_j Y_{ijh} - \rho_j \cdot d_j X_{ijh}$, $d_j Y_{ijh} = Y_{ijh} - Y_{ij,h+1}$, $d_j X_{ijh} = X_{ijh} - X_{ij,h+1}$, $\rho_j = Y_{.j.} / X_{.j.}$, $\zeta_j = [1 - n_j/(10N_j)]\, n_j / [2(n_j - 1)]$, n_j 는 경제활동인구조사에서 j 번째 범주에 대한 표본조사구 수를 나타내며, N_j 는 j 범주에 대한 모집단의 조사구 수를 나타낸다.

3. 복합추정량

i 번째 소지역에 대한 복합추정량 $\hat{Y}_{i.}^{C}$ 는 다음 식을 이용하여 추정할 수 있다.

$$\hat{Y}_{i.}^{C} = \hat{\omega}_{i(opt)} \hat{Y}_{i.} + (1 - \hat{\omega}_{i(opt)}) \hat{Y}_{i.}^{S}, \quad i = 1,2,\cdots,I \quad (11)$$

여기에서 가중값 $\hat{\omega}_{i(opt)} = \dfrac{\hat{Var}(\hat{Y}_{i.}^{S})}{\hat{Var}(\hat{Y}_{i.}^{S}) + \hat{Var}(\hat{Y}_{i.})}$, $i = 1,2,\cdots,I$ 이다.

직접추정량과 합성추정량의 공분산이 $Cov(\widehat{Y_{i.}}, \hat{Y}_{i.}^{S}) = 0$ 라는 가정 하에 복합추정량의 분산추정은 다음 식으로부터 계산될 수 있다.

$$\hat{Var}(\hat{Y}_{i.}^{C}) = \hat{\omega}_{i(opt)}^{2} \hat{Var}(\hat{Y}_{i.}) + (1 - \hat{\omega}_{i(opt)})^{2} \hat{Var}(\hat{Y}_{i.}^{S}) \quad (12)$$

위에서 언급한 직접추정량과 합성추정량 그리고, 복합 추정량에 의해서 조사된 자료를 근거로 계산된 추정값을 기준으로 하여 통계청에서 발표하는 대전광역시의 2007년 6월 실업자 수와 일치시키기 위해서 다음과 같은 비추정식을 이용하여 보정하였다.

$$\hat{Y}_{i.}^{A} = \left(\frac{\hat{Y}_{i.}^{*}}{\sum_{i} \hat{Y}_{i.}^{*}} \right) \hat{Y} \quad (13)$$

〈표 3〉 대전광역시의 구별 실업자 수 추정

(단위: 명)

구	직접추정법			합성추정법			복합추정법		
	추정량	표준오차	CV(%)	추정량	표준오차	CV(%)	추정량	표준오차	CV(%)
동구 (15)	5,746	2,704	47.1	4,477	350	7.81	4,476	348	7.78
중구 (16)	2,460	828	33.6	4,837	234	4.83	4,859	232	4.77
서구 (27)	11,009	2,454	22.3	9,081	883	9.72	9,070	878	9.68
유성구 (9)	1,962	626	31.9	4,167	320	7.69	4,160	315	7.57
대덕구 (14)	5,210	1,703	32.7	3,825	244	6.39	3,822	243	6.37
합계	26,387			26,387			26,387		

$\hat{Y}$는 대전시의 실업자 수의 직접추정량(26,387)이고, $\hat{Y}_i^*$는 소지역 i의 복합 혹은 합성 추정방법의 추정량이다. 〈표 3〉에 구별 실업자 수에 대한 추정량과 각 추정량의 표준오차(S.E)와 상대표준오차 개념인 CV 값을 소지역 추정방법에 따라 정리하였다. 취업자 수와 경제활동인구에 대한 소지역 추정치도 동일한 방법을 적용하여 얻을 수 있다.

〈표 3〉에서 볼 수 있듯이 직접추정법의 경우 표본조사구가 적은 구의 상대표준오차(CV)가 표본조사구가 상대적으로 많은 구에 비해 높고 CV 값이 20% 이상이므로 국가통계로 활용하기에는 문제가 있다. 합성추정량과 복합추정량의 경우에는 CV 값이 10% 미만이다. 5개 구의 합성추정값과 복합추정값의 상대표준오차가 낮고 차이가 크지 않은 이유는 "Borrow Strength"를 근간으로 한 합성추정법의 성질에서 기인한 것이다. 대전 5개 구 소지역 추정 결과 복합추정량이 합성추정량보다 CV 값이 낮고 상대표준오차가 국가통계로 활용할 수 있는 수준인 10% 미만이므로 복합추정법이 실업자 수 추정에 있어서 소지역 추정방법으로 가장 효율적이라 할 수 있다.

4. 잭나이프 방법에 의한 MSE 추정

합성추정량의 추정분산은 식(10)에 의해 계산될 수 있으나 현행 경제활동인구조사 체계에서 편향에 대한 추정은 결코 쉬운 문제가 아니다. 소지역 i 의 실업자 총계에 대한 참값을

센서스 자료를 이용하여 결정할 수는 있으나 시점 상으로 서로 상이한 양상을 보일 가능성이 있기 때문에 편향추정에 직접적으로 이용될 수는 없다.

이 문제의 해결방안으로 Ghosh & Rao(1994)는 $Cov(\widehat{Y}_{i.}, \widehat{Y^*}_i) = 0$이라는 가정 하에서 합성추정량 근사적인 불편추정량을 이용할 것을 제안하였다. 그러나 이 추정량은 소지역에 배정된 표본조사구 수가 충분하지 못할 경우, 직접추정값의 불안정에 기인하여 합성추정값의 평균제곱오차의 추정값이 음의 값이 나올 가능성도 있어 소지역에 배정된 표본조사구 수가 충분하지 못한 한국의 경제활동인구조사 체계에 적용하기에는 무리가 있는 추정공식이다.

이런 문제의 대안으로써 잭나이프 추정방법이 고려될 수 있다. 잭나이프 추정방법의 첫 번째 단계는 경제활동인구조사 자료로부터 반복표본을 생성하는 것이다. 우선 소지역 i 내에서 하나의 표본조사구가 교대로 선택되어 표본으로부터 제거된 후 나머지 표본조사구들에 대해서 승수가 보정된다. 반복표본들은 조사구의 수만큼 생성되며, 이들 반복표본들을 이용하여 새로운 합성추정값들을 다시 계산한다. 잭나이프 추정법을 이용하여 대전광역시의 구별 실업자 추정량의 평균제곱오차를 산정하려면, 한 구에 대한 자료를 제외한 상태에서 다른 4개의 구를 이용한 잭나이프 추정이 아니고 각 구 내에서 조사구 하나씩을 제외하고 나머지 조사구로 해당구의 실업자를 추정하는 방식이 적용되어야 할 것이다. 잭나이프 추정법에 의한 복합추정량의 MSE추정에 대한 구체적인 절차는 아래와 같다.

절차 1. 각 구별로 모든 표본자료를 이용하여 식(5), 식(6)와 식(7)를 이용하여 실업자를 추계하고, 각 구에서 조사구를 하나씩 제외한 후에 동일한 식을 이용하여 구별 실업자를 추정한다.

절차 2. 각 구별로 합성추정량의 MSE 잭나이프 추정값을 다음 식으로 계산한다.

$$mse(\hat{Y}^S_{i.}) = \left\{\widehat{Bias}_{Jack,i}(\hat{Y}^S_{i.})\right\}^2 + \hat{Var}_{Jack,i}(\hat{Y}^S_{i.})$$

$$= (n_i - 1)^2 \left(\frac{\sum_{k=1}^{n_i} \hat{Y}^S_{i.}(k)}{n_i} - \hat{Y}^S_{i.} \right)^2 + \frac{n_i - 1}{n_i} \sum_{k=1}^{n_i} \left(\hat{Y}^S_{i.}(k) - \frac{\sum_{k=1}^{n_i} \hat{Y}^S_{i.}(k)}{n_i} \right)^2 \qquad (14)$$

절차 3. 절차 2에서 계산한 결과를 이용하여 복합추정량에 대입할 가중값을 계산한다.

$$\hat{\omega}_{i(opt)} = \frac{mse(\hat{Y}_{i\cdot}^{S})}{mse(\hat{Y}_{i\cdot}^{S}) + \hat{Var}(\hat{Y}_{i\cdot})} \tag{15}$$

절차 4. 구별로 복합추정량의 MSE 잭나이프 추정값을 아래의 식으로 산정한다.

$$mse(\hat{Y}_{i\cdot}^{C}) = \left\{\widehat{Bias}_{Jack,i}(\hat{Y}_{i\cdot})\right\}^2 + \hat{Var}_{Jack,i}(\hat{Y}_{i\cdot}^{C})$$

$$= (n_i - 1)^2 \left(\frac{\sum_{k=1}^{n_i} \hat{Y}_{i\cdot}^{C}(k)}{n_i} - \hat{Y}_{i\cdot}^{C} \right)^2 + \frac{n_i - 1}{n_i} \sum_{k=1}^{n_i} \left(\hat{Y}_{i\cdot}^{C}(k) - \frac{\sum_{k=1}^{n_i} \hat{Y}_{i\cdot}^{C}(k)}{n_i} \right)^2 \tag{16}$$

절차 5. 절차 3과 절차 4에 의해 계산된 MSE 추정값을 이용하여 구별 실업자의 추정값을 식(8)에 대입하여 최종적으로 실업자를 추계한다.

위의 절차에 따라서 계산한 결과들이 다음 〈표 4〉에 주어졌다.

〈표 4〉에서 볼 수 있듯이 잭나이프 평균제곱오차의 추정값이 가장 낮음을 확인할 수 있다. 이는 대전광역시에서는 모든 구에서 성별 - 연령대별 실업률이 유사하다는 합성추정량에서 전제조건이 크게 틀리지 않음을 보여주고 있다고 생각된다. 합성추정량에서 "Borrow Strength"의 전제조건으로 가정한 내용이 크게 틀리지 않은 경우에는 복합추정량의 분산이 가장 작기 때문에 소지역 추정법에서 바람직한 추정량으로 복합추정량을 적용하는 것이 바람직하다.

〈표 4〉 MSE의 잭나이프 추정값

구	실업자	합성 추정분산	복합 추정분산	잭나이프 MSE
동구	4,965	77,198	77,254	76,894
중구	1,903	34,503	34,531	34,140
서구	9,600	492,469	492,513	488,535
유성구	1,526	64,960	64,981	62,841
대덕구	4,799	37,525	37,727	37,622
합 계	22,793			

IV. 결론

많은 분야에서 소지역 추정법을 적용하여 통계를 생산해야 할 필요성은 대두되고 있으나 이론적 연구뿐만 아니라 보조정보로 이용할 행정업무 자료들도 미흡하기 때문에 앞으로 많은 연구가 진척되어야 할 것이다. 본 연구에서는 소지역 추정법을 개략적으로 요약하여 소지역 추정법을 연구하고 적용하고자 할 때 도움이 되도록 하였다. 특히 소지역 추정법을 직접추정법과 간접추정법 및 모형기반 추정기법으로 세분하여 이용자의 수준에 적합한 추정법을 응용할 수 있게 하였다. 설계기반 소지역 추정법에서 유용한 합성 추정법과 복합추정법을 적용하는 데 가장 큰 애로사항은 평균제곱오차를 정확하게 추정할 수 없다는 데에 있다. 소지역 추정량들의 평균제곱오차에 대한 추정값을 계산할 때 정규성이나 분포의 대칭성을 전제조건으로 가정하였으나 현실적으로 가정이 성립할 수 없고, 성립한다고 하더라도 간단하게 계산할 수 있는 명확한 형태의 공식을 만들 수 없다. 이와 같은 여건에서 복잡한 추정과정을 통해서 계산해야 하는 추정오차 또는 MSE를 간단한 알고리즘을 통해서 계산할 수 있는 잭나이프 추정법을 연구하였다. 알고리즘의 원리는 간편하지만 계산과정이 어렵고 컴퓨터를 통해서만 수행되어야 하는 어려운 점이 있기도 하다.

실제로, 경제활동인구조사의 자료를 이용하여 시·군·구별 실업자를 추정하는 데 세 가지 추정법을 적용하여 복합추정량이 적합함을 입증하였다. 대전광역시의 2007년 6월 경제활동인조사 자료를 기반으로 5개 구별 실업자를 추정해 본 결과, 직접추정량은 표본조사구가 적은 경우 상대표준오차가 30%이상이나 되어 매우 불안정하나 합성 추정량과 복합 추정량은 상대적으로 안정적이고 정확성이 향상된 추정값을 제공하였다.

합성추정량과 복합추정량의 MSE를 좀더 정확하게 계산하기 위해 잭나이프 추정법을 적용한 결과, 표본조사구의 수가 어느 정도로 큰 구에서는 본 연구에서 소개한 근사적 계산법으로 추정한 것과 대동소이함을 보여 주었다. 잭나이프 추정법을 적용하여 복합추정량의 MSE 추정값을 계산할 수 있는 효율적인 알고리즘과 프로그램이 개발된다면, 시·군·구별 실업통계를 작성할 때 소지역 추정법을 적용함으로써 조사구를 증편하지 않고도 일정 수준의 정확도를 갖는 통계를 생산할 수 있으므로 경제적으로 많은 이득을 얻을 수 있다.

본 연구에서 대전광역시의 사례를 연구하였으나 다른 광역시나 도지역에서도 유사한 결과를 얻을 수 있을 것으로 기대한다. 시간과 예산을 투입하여 심층적으로 연구한다면 소지역 추정법을 적용하여 우리나라 시·군·구 단위의 실업자를 추계할 수 있을 것이다. 뿐만 아니라 본 연구를 확대 적용한다면 실업통계뿐 아니라, 보건과 건강관련 통계에서도 시·도

단위까지 일정 수준의 정확도를 갖춘 통계를 생산하는 데 있어서 표본 크기를 늘리지 않고도 가능할 것으로 생각된다. 앞으로 정보화가 심화되면 될수록 통계생산에서는 더 정확성이 요구될 것이며, 또한 통계 생산단위도 더욱 세분화될 것이므로 소지역 추정법의 연구는 더 필요할 것이다.

참고문헌

김수택·고석남·김상대. 2008. "소지역 노동통계의 효율적 추정방안 -실업률을 중심으로-." 《산업관계연구》 18(1): 53-76.

이계오. 2000. "시군구 실업자 추정을 위한 소지역 추정법." 《응용통계연구》 13(2): 275-286.

정연수·이계오·이우일. 2003. "시군구 실업자 총계 추정을 위한 설계기반 간접추정법." 《응용통계연구》 16(1): 1-14.

통계청. 2000. 《소지역 추정법 연구》 통계청 조사관리과.

Drew, J.D., Singh, M.P., and Choudhry, G.H. 1982. "Evaluation of Small Area Estimation Techniques for the Canadian Labour Force Survey." *Survey Methodology* 8: 17-47.

Ghosh, M. and Rao, J. N. K. 1994. "Small Area Estimation: an Appraisal." *Statistical Science* 9: 55-93.

Jiang, J. and Lahiri, P. 2002. "A Unified Jackknife Theory for Eempirical Best Prediction with M-estimation." *Annals of Statistics* 30(6): 1782-1810.

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

Sandal, C. E. 1984. "Design-consistent Versus Model-dependent Estimation for Small Domains." *Journal of the American Statistical Association* 79: 624-631.

Statistics Canada. 1998. *Guide to the Labour Force Survey*. Available on the internet at www.statcan.ca/english/concepts/labour/index.htm

[접수 2009/1/28, 수정 2009/2/10, 게재확정 2009/2/20]