응용통계연구 제 **16**권 *1*호, *2003*년, *pp. 1-14*

시군구 실업자 총계 추정을 위한 설계기반 간접추정법

정연수 [1] 이계오 [2] 이우일 [3]

요 약

본 연구에서는 현행 경제활동인구조사 체계에 근거하여 대영역 내의 시군구 단위 행정자치구역들에 대한 실업통계들을 생산할 수 있는 소지역 추정법이 제안된다. 고려된 소지역 추정량들은 합성추정량, 복합추정량과 같은 설계기반 간접추정량들이며 이러한 추정량들에 대한 평균제곱오차 추정식이 경제활동인구조사 체계 하에서 산정되어 시군구 단위 소지역 추정값들에 대한 정확도의 측도로써 활용된다. 2000년 12월 충북지역의 경제활동인구조사 자료로부터 이 지역 내의 10개 시군구 단위 행정자치구역들에 대한 실업자 총계 및 잭나이프 평균제곱오차가 본 연구에서 제시된 추정절차에 의해 추정된다. 시군구 단위 실업자 총계 추정값들의 신뢰성은 이들 추정값들의 상대편향(Relative Bias)과 상대오차제곱근(Relative Root Mean Square Error)에 의해 평가된다. 현행 한국 경제활동인구조사 체계 하에서 복합추정량이 다른 추정량들에 비해 매우 안정적임을 밝힌다.

주요용어: 소지역추정, 합성추정량, 복합추정량, 잭나이프 평균제곱오차, 상대편향, 상대오차제곱근

1. 서 론

통계청에서는 취업, 실업 등과 같은 경제활동 특성을 조사하여 국가 고용정책 입안과 평가에 필요한 기초자료를 수집할 목적으로 매월 약 30,000 표본가구 내에 거주하는 만 15세 이상인 사람들을 대상으로 경제활동인구조사를 실시하고 있다. 매월 15일을 포함하는 주중에 표본가구 내에 거주하는 조사 대상자들에 대해 취업, 실업 및 비경제활동인구 관련 사항들이 방문 면접이나 컴퓨터 면접 방식으로 조사되며, 조사된 자료는 직접추정방법에 의해 세부 항목별 통계가 산정되어, 매 익월 말경 대영역인 7개 광역시와 9개 도 단위에 대해 공표된다.

1995년 지방자치제 출범 이후 약 7년이 경과되기까지 많은 시군구 단위 행정자치구역들이 시군구 단위 소지역 통계 생산을 요구하고 있으나 현재와 같은 대영역 표본설계를 기반으로 하는 통계 작성방법으로는 신뢰성있는 소지역 통계 생산은 불가능한 실정이다. 소지역 통계 생산은 단순히 추정단계에서 추정량의 선택을 통해서 해결될 수 있는 것이 아니

1) (363-849) 충북 청원군 남일면 쌍수리 사서함 335-2, 공군사관학교 전산통계학과, 부교수
E-mail: yschung@afa.ac.kr
2) (110-054) 서울 종로구 사직동 208, 한국갤럽 조사연구소 자문교수
E-mail: kayolee@orgio.net
3) (363-849) 충북 청원군 남일면 쌍수리 사서함 335-2, 공군사관학교 전산통계학과, 조교수
E-mail: wilee@afa.ac.kr

라 통계조사의 계획, 표본설계, 추정 등 통계조사 전과정을 종합적으로 고려할 때 가능한 일이기 때문이다.

통계청의 경제활동인구조사는 대영역인 광역시와 도 단위의 통계 작성을 목적으로 표본설계 되었기 때문에 소지역인 시군구 단위 행정자치구역들은 표본설계에 반영된 관심영역들이 아니다. 따라서 현재 활용되고 있는 대영역 기반의 표본설계를 이용하여 시군구 단위 소지역 통계를 직접 생산할 경우, 시군구 지역에 배정된 표본조사구 수가 매우 불균형적이고 특정 시군구에 대해서는 하나 내지 둘 정도의 작은 표본조사구 수가 배정되어 있기 때문에 신뢰할만한 소지역 통계 생산은 어렵게 된다.

우리의 목적은 대영역 기반 표본설계 구조 하에서 직접 산정된 시군구 단위 직접추정값들을 명시적인 모형을 통해 보정하여 추정값들의 신뢰성을 확보하는 일이다. 이 논문에서는 현재 통계청에서 실시하고 있는 대영역 기반 경제활동인구조사 체계 하에서 산정된 시군구 단위의 소지역들에 대한 직접추정값들을 본 논문에서 제안한 합성추정법 또는 복합추정법과 같은 설계기반 간접추정법을 통해 보정하여 신뢰성있는 시군구 단위들에 대한 실업자 총계를 추론한다.

현 경제활동인구조사 체계 하에서 산정된 시군구 단위 월별 직접추정값들은 표본조사구 수가 불균형적으로 할당된 상태에서 추정된 매우 불안정한 추정값들이므로 센서스 및 행정보고자료를 통해 선택된 다음과 같은 보조정보들을 이용하여 보정을 시도하였다. 시군구 단위 소지역들과 유사한 특성을 갖는 인근 유사지역의 정보를 시군구 단위 소지역 추정에 적용하기 위해서 우선 경제활동인구조사 자료를 대영역별로 크게 시군구 그룹으로 분할하고, 각 그룹 내에서는 유사성질 범주를 성별(남,여)과 연령대별(15-34세, 35세 이상)로 총 4개의 범주로 구분하였다. 각 그룹에서는 성별에 따른 경제활동참가율을 그리고 각 그룹별 4개 범주에 대해서는 범주별 실업률을 산출하여 대영역 내의 시군구 단위들에 대한 실업자 총계 추정을 위한 보조정보로 활용하였다.

2. 직접추정법

소지역 i 의 실업자 총계에 대한 직접추정값은 소지역에 배정된 표본만을 이용하여 추정되며, 경제활동인구조사 체계에서의 실업자 총계에 대한 직접추정공식은 다음과 같이 주어진다.

$$\begin{aligned} \widehat{Y}_{i\cdot} &= \sum_{s=1}^{2} {}_s\widehat{Y}_{i\cdot} \quad , \; i = 1, \ldots, I; s = 1, 2; h = 1, \ldots, n_i \\ &= \sum_{s=1}^{2} \sum_{h=1}^{n_i} {}_s\widehat{Y}_{ih} \\ &= \sum_{s=1}^{2} \sum_{h=1}^{n_i} {}_sM_i \; {}_sY_{ih} \; , \end{aligned} \tag{2.1}$$

여기에서 s는 성별(남,여)을 나타내며, n_i 는 경제활동인구조사에서 소지역 i 에 할당된 표본조사구 수, ${}_sY_{ih}$는 각 성별에 대해서 소지역 i 의 h 번째 표본조사구에서 조사된 실업자

수를 나타낸다. 승수 ${}_sM_i = {}_s\widehat{X}_i/{}_sX_{i\cdot}$는 직접추정량 $\widehat{Y}_{i\cdot}$이 불편추정량이 되도록 산정된다. 승수의 표현식에서 ${}_s\widehat{X}_i$ 는 소지역 i 에 대한 15세 이상의 상주추계인구를 나타내며, ${}_sX_{i\cdot}$는 경제활동인구조사에서 집계된 15세 이상의 상주조사인구를 나타낸다.

소지역 i에 대한 직접추정량 $\widehat{Y}_{i\cdot}$의 분산은 다음과 같이 주어진다.

$$\begin{aligned} Var(\widehat{Y}_{i\cdot}) &= \sum_{s=1}^{2} Var({}_s\widehat{Y}_{i\cdot}) + 2Cov({}_1\widehat{Y}_{i\cdot}, {}_2\widehat{Y}_{i\cdot}) \\ &= \sum_{s=1}^{2} {}_sM_i^2 Var(\sum_{h=1}^{n_i} {}_sY_{ih}) + 2\,{}_1M_i\ {}_2M_i Cov(\sum_{h=1}^{n_i} {}_1Y_{ih}, \sum_{h=1}^{n_i} {}_2Y_{ih}) \end{aligned} \tag{2.2}$$

식(2.2)는 경제활동인구조사 자료로부터 추정되어야 하며, 통계청에서는 연속차 분산추정식을 적용하여 직접추정량들에 대한 분산추정값들을 산정한다. 연속차 분산추정식은 다음 식(2.3)과 같이 주어진다.

$$\widehat{Var}(\widehat{Y}_{i\cdot}) = \sum_{s=1}^{2} {}_sM_i^2(\xi_i \sum_{h=1}^{n_i} {}_sU_{ih}^2) + 2\ {}_1M_i\ {}_2M_i(\xi_i \sum_{h=1}^{n_i} {}_1U_{ih}\ {}_2U_{ih}), \tag{2.3}$$

여기에서

$$\begin{aligned} {}_sU_{ih} &= d\ {}_sY_{ih} - {}_s\rho_i \cdot d\ {}_sX_{ih}, \\ d\ {}_sY_{ih} &= {}_sY_{ih} - {}_sY_{i,h+1}, \\ d\ {}_sX_{ih} &= {}_sX_{ih} - {}_sX_{i,h+1}, \\ {}_s\rho_i &= {}_sY_{i\cdot}/{}_sX_{i\cdot}, \\ \xi_i &= [(1-f_i)n_i]/[2(n_i-1)], \\ f_i &= n_i/(10N_i) \end{aligned}$$

이며, N_i는 표본추출틀에서 소지역 i의 조사구 수를 나타낸다.

경제활동인구조사에서 대영역에 포함된 시군구 단위 소지역들은 표본설계에 반영된 관심영역들이 아니다. 따라서 대영역 표본설계에 기반을 둔 표본조사로부터 시군구 단위 소지역들에 대한 실업자 총계를 산정한다면, 시군구 단위에 배정된 표본조사구의 수가 충분하지 않기 때문에 신뢰할 만한 추정결과를 얻을 수 없게 된다. 이러한 관점에서 현행 경제활동인구조사 자료로부터 시군구 단위 소지역 추정값들의 신뢰성을 확보하기 위한 합성추정법 및 복합추정법과 같은 설계기반 간접추정법들이 제안될 수 있다.

3. 합성추정법

3.1. 합성추정량

대영역 표본설계에 기반을 둔 통계청의 직접추정량은 각 소지역에 할당된 표본조사구 수가 충분하지 못할 경우 소지역 단위의 실업통계에 대한 정확도를 제공하지는 못한다. 합

성추정량은 추정 대상 소지역과 특성이 유사한 인근 소지역들의 정보를 추정에 이용하는 간접적인 설계기반 추정량이다. 여기에서 해당 소지역 추정 시 이용되는 인근 소지역들에 대한 정보를 총칭하여 "Borrow Strength"라 부른다.

우선 대영역 내에 I개의 소지역들이 있다고 가정하자. "Borrow Strength"를 적용하기 위해 대영역을 특성이 유사한 K개의 부차 관심영역들로 분할하고, 각 부차 관심영역들을 J개의 성별-연령대별 범주로 구분한다. 여기에서 $I = \sum_{k=1}^{K} I_k$이며, 분할된 부차 관심영역들은 각각 I_k개의 동질적인 소지역 단위들로 구성된다.

"Borrow Strength"를 적용하여 정의되는 합성추정량은 해당 소지역과 유사한 정보를 갖는 인근 소지역들의 정보를 이용하여 추정되므로 추정오차는 직접추정량에 비해 현저하게 줄어들 수 있으나 해당 소지역과 인근 유사지역의 정보가 동질적이지 못할 경우 편향이 발생할 가능성이 있다.

합성추정량을 정의하기 위해 다음과 같은 기호들이 이용되었다.

N_i = 표본추출틀에서 소지역 i의 조사구 수,
n_i = 경제활동인구조사에서 소지역 i에 할당된 표본조사구 수,
${}_jP^C_{1995,i}$ = 1995년 센서스로부터 추계된 j범주에 대한 소지역 i 의 상주인구,
${}_jP^R_{1995,i}$ = j범주에 대한 소지역 i의 1995년 주민등록인구,
${}_jP^R_{month,i}$ = j범주에 대한 소지역 i의 경제활동인구조사 달의 주민등록인구,
${}_j\widehat{X}_i$ = j범주에 대한 소지역 i의 상주추정인구,
${}_jY_{ih}$ = j범주에 대한 소지역 i의 h번째 표본조사구의 실업자 수.

I_k개의 소지역들을 포함하고 있는 각 부차 관심영역 내에서 소지역 i의 실업자 총계에 대한 합성추정량 $\widehat{Y}^S_{i\cdot}$는 다음과 같이 정의될 수 있다.

$$\widehat{Y}^S_{i\cdot} = \sum_{j=1}^{J} \frac{{}_j\widehat{P}_i}{{}_j\widehat{X}_{\cdot}} {}_j\widehat{Y}_{dir} \ , \ i = 1, \ldots, I_k \ , \tag{3.1}$$

여기에서

$$\begin{aligned} {}_j\widehat{P}_i &= \frac{{}_jP^C_{1995,i}}{{}_jP^R_{1995,i}} {}_jP^R_{month,i} \ , \\ {}_j\widehat{X}_{\cdot} &= \sum_{i=1}^{I_k} {}_j\widehat{X}_i \ , \\ {}_j\widehat{Y}_{dir} &= \sum_{i=1}^{I_k} \sum_{h=1}^{n_i} {}_jM_i \ {}_jY_{ih} \end{aligned}$$

로 주어진다. ${}_j\widehat{P}_i$는 행정보고자료로부터 산정된 j범주에 대한 소지역 i의 상주추정인구를 나타내며, ${}_j\widehat{X}_{\cdot}$는 경제활동인구조사 자료로부터 산정되는 j범주에 대한 상주추정인구를 나타낸다. 또한 ${}_j\widehat{Y}_{dir}$는 j번째 성별-연령대별 범주의 실업자 총계에 대한 직접추정량을 의미하며, 경제활동인구조사 자료로부터 산정된다. 소지역 i의 j범주에 대한 승수는 ${}_jM_i = {}_j\widehat{X}_i / {}_jX_{i\cdot}$로 주어진다.

3.2. 평균제곱오차 추정

합성추정량 $\widehat{Y}_{i\cdot}^{S}$에 대한 정확도의 측도로써 다음과 같은 평균제곱오차(MSE)를 고려할 수 있다.

$$MSE(\widehat{Y}_{i\cdot}^{S}) = Var(\widehat{Y}_{i\cdot}^{S}) + [Bias(\widehat{Y}_{i\cdot}^{S})]^2 \ . \tag{3.2}$$

식(3.2)에서 $Var(\widehat{Y}_{i\cdot}^{S})$는 ${}_j\widehat{P}_i/{}_j\widehat{X}_{\cdot} = const$ 를 가정한다면 다음 식과 같이 주어질 수 있다.

$$\begin{aligned} Var(\widehat{Y}_{i\cdot}^{S}) &= \sum_{j=1}^{J}\left(\frac{{}_j\widehat{P}_i}{{}_j\widehat{X}_{\cdot}}\right)^2 Var({}_j\widehat{Y}_{dir}) \ , \quad i=1,\ldots,I_k \\ &+ 2\sum_{j<l}\left(\frac{{}_j\widehat{P}_i}{{}_j\widehat{X}_{\cdot}}\right)\left(\frac{{}_l\widehat{P}_i}{{}_l\widehat{X}_{\cdot}}\right) Cov({}_j\widehat{Y}_{dir}, {}_l\widehat{Y}_{dir}). \end{aligned}$$

j범주의 실업자 총계에 대한 직접추정량은

$${}_j\widehat{Y}_{dir} = \sum_{i=1}^{I_k}\sum_{h=1}^{n_{ij}} M_j Y_{ijh}$$

로 주어지므로 직접추정량 ${}_j\widehat{Y}_{dir}$의 분산과 공분산은 연속차 분산추정방법에 의해 다음 식으로부터 추정될 수 있다.

$$\begin{aligned} \widehat{Var}({}_j\widehat{Y}_{dir}) &= M_j^2\xi_j\sum_{i=1}^{I_k}\sum_{h=1}^{n_{ij}} U_{ijh}^2 \ , \\ \widehat{Cov}({}_j\widehat{Y}_{dir}, {}_l\widehat{Y}_{dir}) &= M_jM_l\xi_j\sum_{i=1}^{I_k}\sum_{h=1}^{n_{ij}} U_{ijh}U_{ilh}, \end{aligned}$$

여기에서

$$\begin{aligned} U_{ijh} &= d\ {}_jY_{ijh} - \frac{Y_{\cdot j\cdot}}{X_{\cdot j\cdot}}\cdot d\ {}_jX_{ijh} \ , \\ d\ {}_jY_{ijh} &= Y_{ijh} - Y_{ij,h+1} \ , \\ d\ {}_jX_{ijh} &= X_{ijh} - X_{ij,h+1} \ , \\ \xi_j &= \frac{(1-f_j)n_j}{2(n_j-1)} \ , \\ f_j &= \frac{n_j}{10N_j} \end{aligned}$$

로 주어진다.

따라서 합성추정량 $\widehat{Y}_{i\cdot}^{S}$의 추정분산은 다음과 같이 주어진다.

$$\begin{aligned} \widehat{Var}(\widehat{Y}_{i\cdot}^{S}) &= \left(\frac{{}_j\widehat{P}_i}{{}_j\widehat{X}_{\cdot}}\right)^2\left(M_j^2\xi_j\sum_{i=1}^{I_k}\sum_{h=1}^{n_{ij}} U_{ijh}^2\right) \\ &+ 2\sum_{j<l}\left(\frac{{}_j\widehat{P}_i}{{}_j\widehat{X}_{\cdot}}\right)\left(\frac{{}_l\widehat{P}_i}{{}_l\widehat{X}_{\cdot}}\right)\left(M_jM_l\xi_j\sum_{i=1}^{I_k}\sum_{h=1}^{n_{ij}} U_{ijh}U_{ilh}\right). \end{aligned}$$

소지역 i에 대한 합성추정량 $\widehat{Y}_{i\cdot}^S$의 추정분산은 앞에서 설명된 것과 같이 명시적인 절차에 의해 산정될 수는 있으나, 현행 경제활동인구조사 체계에서 편향에 대한 추정은 결코 쉬운 문제가 아니다. 소지역 i의 실업자 총계에 대한 참값을 1995년 센서스자료를 이용하여 결정할 수는 있으나 시점 상으로 서로 상이한 양상을 보일 가능성이 있기 때문에 편향 추정에 직접적으로 이용될 수는 없다. 이러한 문제에 기인하여 Ghosh and Rao (1994)는 $Cov(\widehat{Y}_{i\cdot}, \widehat{Y}_{i\cdot}^S) = 0$ 의 가정 하에서 다음과 같은 근사적인 MSE 추정량을 이용할 것을 제안한 바 있다.

$$mse(\widehat{Y}_{i\cdot}^S) \approx (\widehat{Y}_{i\cdot}^S - \widehat{Y}_{i\cdot})^2 - \widehat{V}ar(\widehat{Y}_{i\cdot}). \tag{3.3}$$

식(3.3)은 $MSE(\widehat{Y}_{i\cdot}^S)$에 대한 근사적인 불편추정량이나 소지역 i에 할당된 표본조사구 수가 충분하지 못할 경우 직접추정값의 추정오차가 커져 MSE 추정값이 음의 값이 나올 가능성도 있다. 따라서 소지역에 할당된 표본조사구 수가 충분하지 않은 우리나라의 경제활동인구조사 체계에 적용하기에는 무리가 따르는 추정공식이다. 또 다른 방법으로 부차 관심영역 내의 모든 소지역들에 대해 MSE 추정값들의 평균을 취하는 방법이 이용될 수 있으나 이 측도는 해당 소지역들에 대한 MSE 추정값들을 제공하지는 못한다.

MSE 추정을 위한 하나의 대안으로써 잭나이프 추정방법이 고려될 수 있다. 잭나이프 추정에서는 서로 독립이고 동일한 임의 확률분포를 따르는 확률표본을 가정한다. 우리나라의 경제활동인구조사 표본은 2단 층화추출로 추출되는 확률표본이다. 대영역 내에서 표본조사구들이 일차적으로 계통추출된 후 각 표본조사구 내에서 일정 량의 가구 수를 갖는 작은 구획들이 이차적으로 랜덤추출되어 구획 내의 가구단위들에 대해 조사가 이루어진다. 표본조사구 내의 구획단위들은 서로 독립인 확률표본이고 이들의 관측값에 의해 표본조사구의 관측값이 결정된다. 그러나 대영역 내의 표본조사구들은 랜덤 수에 의해 계통추출되므로 유사 독립표본으로 간주될 수는 있지만 정확히 독립인 확률표본들은 아니다. 현재 캐나다 통계청에서는 실업통계에 대한 분산 계산 시 대영역 내에서 계통추출된 표본조사구들을 유사 독립표본으로 가정하고 잭나이프 추정방법을 적용하고 있다 (Statistics Canada, 1998).

대영역 내의 표본조사구들을 서로 독립이고 동일한 임의 분포를 갖는 확률표본으로 가정한 상태에서 잭나이프 추정절차를 소개하면 다음과 같다. 잭나이프 추정에서 첫번째 단계는 경제활동인구조사 자료로부터 반복표본을 생성하는 일이다. 우선 소지역 i 내에서 하나의 표본조사구가 교대로 선택되어 표본으로부터 제거된 후 나머지 표본조사구들에 대해서 승수가 보정된다. 반복표본들은 표본조사구의 갯수만큼 생성되며, 이들 반복표본들을 이용하여 새로운 합성추정값들이 다시 계산된다. 구체적인 절차를 명시하면 다음과 같다.

(i) 대영역 내에서 분할된 k번째 부차 관심영역 내에서 소지역 i로부터 h번째 표본조사구를 제거한 후, 실업자 수 Y에 대한 다음과 같은 반복표본을 생성한다.

$$S_{i(h)} = \left\{ Y_{11}, \ldots, Y_{1n_1}; \ldots; Y_{i1}, \ldots, Y_{i,h-1}, Y_{i,h+1}, \ldots, Y_{in_i}; \ldots; Y_{I_k1}, \ldots, Y_{I_kn_{I_k}} \right\}.$$

기호 $i(h)$는 새로운 합성추정값을 생성하기 위해 소지역 i로부터 h번째 표본조사구가 제거되었다는 것을 나타낸다. 소지역 i에 대해서 총 n_i개의 반복표본이 생성되며, k번째 부차 관심영역에 대한 반복표본의 수는 총 $n = \sum_{i=1}^{I_k} n_i$개이다.

(ii) 해당 소지역 i에 대해서 소지역 내에 남아있는 $n_i - 1$개의 표본조사구의 전체 조사 가구들에 대해서 승수에 대한 보정이 이루어진다. 보정된 승수값은 다음과 같다.

$$ {}_jM_i^{adj} = \frac{n_i}{n_i - 1} {}_jM_i $$

(iii) k번째 부차 관심영역 내에 남아있는 $n-1$개 표본조사구들을 이용하여 소지역 i에 대한 새로운 합성추정값 $\widehat{Y}_{i\cdot}^S(h)$를 계산한다.

위의 절차는 해당 부차 관심영역 내에 있는 모든 표본조사구들에 대해서 반복되며, 이로부터 n개의 서로 다른 실업자 총계에 대한 합성추정값들이 생성된다. 소지역 i에 대해서는 n_i개의 서로 다른 합성추정값들이 얻어진다.

소지역 i의 실업자 총계에 대한 합성추정값들의 잭나이프 MSE 추정식은 다음과 같이 주어진다.

$$ mse_J(\widehat{Y}_{i\cdot}^S) = \widehat{V}ar_J(\widehat{Y}_{i\cdot}^S) + [\widehat{B}ias_J(\widehat{Y}_{i\cdot}^S)]^2 , \tag{3.4} $$

여기에서

$$ \begin{aligned} \widehat{V}ar_J(\widehat{Y}_{i\cdot}^S) &= \frac{n_i - 1}{n_i} \sum_{h=1}^{n_i} \left[\widehat{Y}_{i\cdot}^S(h) - \frac{1}{n_i} \sum_{l=1}^{n_i} \widehat{Y}_{i\cdot}^S(l) \right]^2 , \\ \widehat{B}ias_J(\widehat{Y}_{i\cdot}^S) &= (n_i - 1) \left[\frac{1}{n_i} \sum_{h=1}^{n_i} \widehat{Y}_{i\cdot}^S(h) - \widehat{Y}_{i\cdot}^S \right] \end{aligned} $$

로 주어진다.

4. 복합추정법

경제활동인구조사 자료로부터 추정된 소지역 i에 대한 직접추정량 $\widehat{Y}_{i\cdot}$는 해당 소지역에 할당된 표본조사구 수가 충분하지 않으므로 추정값들의 신뢰성을 확보할 수 없고, 또한 인근 지역의 유사정보를 이용하여 추정되는 합성추정량 $\widehat{Y}_{i\cdot}^S$는 편향이 내재되어 있을 가능성이 있다. 따라서 소지역에 할당된 표본조사구의 수가 적을 경우, 직접추정량의 불안정성과 합성추정량의 편향 가능성을 보완하기 위해 직접추정량 $\widehat{Y}_{i\cdot}$와 합성추정량 $\widehat{Y}_{i\cdot}^S$의 가중평균을 이용한 다음과 같은 복합추정량 $\widehat{Y}_{i\cdot}^C$가 고려될 수 있다.

$$ \widehat{Y}_{i\cdot}^C = \omega_i \widehat{Y}_{i\cdot} + (1 - \omega_i) \widehat{Y}_{i\cdot}^S , \ i = 1, \ldots, I_k , \tag{4.1} $$

여기에서 가중치 ω_i는 0과 1사이의 값을 취한다.

이때 복합추정량 $\widehat{Y}_{i\cdot}^C$의 MSE는 다음과 같이 주어진다.

$$ \begin{aligned} MSE(\widehat{Y}_{i\cdot}^C) &= \omega_i^2 MSE(\widehat{Y}_{i\cdot}) + (1 - \omega_i)^2 MSE(\widehat{Y}_{i\cdot}^S) \\ &+ 2\omega_i(1 - \omega_i) E(\widehat{Y}_{i\cdot} - Y_{i\cdot}^*)(\widehat{Y}_{i\cdot}^S - Y_{i\cdot}^*) , \end{aligned} \tag{4.2} $$

여기에서 $Y_{i\cdot}^*$는 소지역 i의 실업자 총계에 대한 참값을 나타낸다. 식(4.2)를 ω_i의 함수로 가정하여 가중치 ω_i에 대해서 미분하면, MSE를 최소화하는 다음과 같은 가중치를 산정할 수 있다.

$$\omega_{i(opt)}^* = \frac{MSE(\widehat{Y}_{i\cdot}^S) - E(\widehat{Y}_{i\cdot} - Y_{i\cdot}^*)(\widehat{Y}_{i\cdot}^S - Y_{i\cdot}^*)}{MSE(\widehat{Y}_{i\cdot}^S) + MSE(\widehat{Y}_{i\cdot}) - 2E(\widehat{Y}_{i\cdot} - Y_{i\cdot}^*)(\widehat{Y}_{i\cdot}^S - Y_{i\cdot}^*)} , \tag{4.3}$$

여기에서 직접추정량 $\widehat{Y}_{i\cdot}$는 불편추정량이 되도록 산정하므로 $MSE(\widehat{Y}_{i\cdot})$는 $Var(\widehat{Y}_{i\cdot})$과 동일한 값을 갖는다. 또한 식(4.3)에서 $Cov(\widehat{Y}_{i\cdot}, \widehat{Y}_{i\cdot}^S) = 0$ 를 가정한다면, 가중치 $\omega_{i(opt)}^*$는 다음 식으로 근사될 수 있다.

$$\omega_{i(opt)} = \frac{MSE(\widehat{Y}_{i\cdot}^S)}{MSE(\widehat{Y}_{i\cdot}^S) + Var(\widehat{Y}_{i\cdot})} .$$

위의 $\omega_{i(opt)}$ 은 경제활동인구조사 자료로부터 추정되어야 할 값이다. $MSE(\widehat{Y}_{i\cdot}^S)$ 에 대해서는 식(3.4)에서 주어진 $mse_J(\widehat{Y}_{i\cdot}^S)$으로, $Var(\widehat{Y}_{i\cdot})$에 대해서는 식(2.3)에서 주어진 $\widehat{Var}(\widehat{Y}_{i\cdot})$으로 추정될 수 있고, 이때 최적 가중치에 대한 추정식은 다음과 같이 주어진다.

$$\widehat{\omega}_{i(opt)} = \frac{mse_J(\widehat{Y}_{i\cdot}^S)}{mse_J(\widehat{Y}_{i\cdot}^S) + \widehat{Var}(\widehat{Y}_{i\cdot})} .$$

따라서 식(4.1)의 복합추정량은 경제활동인구조사 자료로부터 추정된 최적가중치 $\widehat{\omega}_{i(opt)}$를 이용하여 다음 식으로부터 추정될 수 있다.

$$\widehat{Y}_{i\cdot}^C = \widehat{\omega}_{i(opt)}\widehat{Y}_{i\cdot} + (1-\omega_i)\widehat{Y}_{i\cdot}^S . \tag{4.4}$$

복합추정량의 MSE추정에서도 합성추정량의 경우와 같이 잭나이프 추정 방법이 이용될 수 있으며, 소지역 i 에서 실업자 총계에 대한 복합추정량의 MSE 추정값들은 다음 추정식으로부터 산정할 수 있다.

$$mse_J(\widehat{Y}_{i\cdot}^C) = \widehat{Var}(\widehat{Y}_{i\cdot}^C) + [\widehat{Bias}_J(\widehat{Y}_{i\cdot}^C)]^2 ,$$

여기에서

$$\begin{aligned}
\widehat{Var}_J(\widehat{Y}_{i\cdot}^C) &= \frac{n_i - 1}{n_i}\sum_{h=1}^{n_i}\left[\widehat{Y}_{i\cdot}^C(h) - \frac{1}{n_i}\sum_{l=1}^{n_i}\widehat{Y}_{i\cdot}^C(l)\right]^2 , \\
\widehat{Bias}_J(\widehat{Y}_{i\cdot}^C) &= (n_i - 1)\left[\frac{1}{n_i}\sum_{h=1}^{n_i}\widehat{Y}_{i\cdot}^C(h) - \widehat{Y}_{i\cdot}^C\right] , \\
\widehat{Y}_{i\cdot}^C(h) &= \widehat{\omega}_{i(opt)}\widehat{Y}_{i\cdot} + (1-\omega_{i(opt)})\widehat{Y}_{i\cdot}^S(h)
\end{aligned}$$

로 주어진다.

5. 자료 분석

대영역인 충북지역에 대해 조사된 2000년 12월 경제활동인구조사 자료를 이용하여 충북지역 내의 시군구 단위 소지역들에 대한 추정결과를 설명하기로 한다. 충북지역은 시군 단위의 11개 행정자치구역들로 구성되어 있다. 경제활동인구조사에서 충북지역에 할당된 표본조사구 수는 64개이고, 조사가구 수는 약 1,500개이다. 여기에서 표본조사구 수가 한개만이 배정되어있는 진천군은 분석에서 제외하였고, 나머지 10개 시군 단위 행정자치구역들에 대해 실업자 총계 추정을 실시하였다.

현 경제활동인구조사의 표본설계에서 관심영역은 충북지역과 같은 대영역들이기 때문에 시군 단위 소지역들인 각 행정자치구역들은 표본설계에서 계획되지 않은 관심영역에 해당된다. 따라서 이들 시군 단위 행정자치구역들에 배정된 표본조사구들 만을 이용하여 실업자 총계 추정값들을 산출한다면 추정오차는 신뢰할 수 없을 정도로 커질 우려가 있다. 현행 경제활동인구조사 체계에서는 각 시군 단위 행정자치구역들에 배정된 표본조사구들의 수가 충분하지 못하기 때문에 보다 정확한 실업자 총계에 대한 추정값들을 산출하기 위해 해당 소지역과 유사한 특성을 갖는 인근 소지역들의 정보를 소지역 추정을 위한 보조정보로 활용하였다.

"Borrow Strength"를 적용하기 위해 우선 대영역인 충북지역을 서로 특성이 유사한 두개의 부차관심영역들로 분할하였다. 이들 부차관심영역들은 시와 군 그룹들이다. 다음으로 각 부차관심영역을 성별(남, 여)-연령대별(15-34세, 35세 이상)의 4개의 범주들로 구분하였다. 각 부차관심영역에 대한 성별-연령대별 4개 범주에 대한 경제활동인구 및 실업자 총계는 경제활동인구조사 자료로부터 추정하였고, 각 시군 단위 행정자치구역들에 대한 범주별 상주추정인구는 통계청의 추계자료를 이용하였다. 이들 정보들을 시군 단위 행정자치구역들에 대한 실업자 총계 추정을 위한 보조정보로 활용하였다.

2000년 12월 경제활동인구조사 자료를 이용하여 추정된 충북 내의 10개 시군 단위 행정자치구역들에 대한 실업자 총계 추정결과가 다음 표 5.1에 주어졌다.

표 5.1 을 살펴보면, 실업자 총계에 대한 직접추정값들의 추정오차는 10개 행정자치구역 모두에서 크게 나타나며 매우 불안정하다. 반면, 합성추정값 및 복합추정값들의 MSE 추정값들은 직접추정값들의 추정오차에 비해 매우 작고 안정적이다. 세가지 추정량들 중 복합추정값들의 MSE 추정값들이 가장 작고 안정적으로 나타난다.

각 행정자치구역들에 대한 직접추정값들의 신뢰성을 평가하기 위해 직접추정값들의 상대추정오차(Relative Standard Error:RSE)를 계산하였다. 합성추정값들과 복합추정값들에 대해서는 상대편향(Relative Bias:RB)값과 상대오차제곱근(Relative Root Mean Square Error:RRMSE)을 계산하여 추정값들의 신뢰성을 평가하였다. 소지역 i에 대한 실업자 총

표 5.1: 충북지역의 시군 단위 행정자치구역들에 대한 실업자 총계 및 MSE 추정치 (2000년 12월)

Area No.	직접추정법		합성추정법		복합추정법		표본 조사구수
	$\widehat{Y}_{i\cdot}$	Est.se	$\widehat{Y}_{i\cdot}^{S}$	$\sqrt{mse_J}$	$\widehat{Y}_{i\cdot}^{C}$	$\sqrt{mse_J}$	
1	8,517	1,733	7,969	580	8,023	493	22
2	3,949	1,445	2,823	725	3,050	607	11
3	365	390	1,830	110	1,723	101	4
4	503	373	612	234	581	196	2
5	781	676	1,164	169	1,140	158	3
6	1,275	577	1,230	282	1,238	233	3
7	1,032	646	1,459	295	1,384	252	5
8	1,795	893	1,825	346	1,821	306	6
9	1,023	602	2,888	574	2,000	270	5
10	512	384	872	94	851	92	2

계 추정량을 $\widehat{Y}_{i\cdot}^{*}$로 나타낼때, RB, RSE 와 RRMSE 는 각각 다음과 같이 주어진다.

$$RB(\widehat{Y}_{i\cdot}^{*}) = \frac{\widehat{Bias}(\widehat{Y}_{i\cdot}^{*})}{\widehat{Y}_{i\cdot}^{*}} \times 100\,,$$

$$RSE(\widehat{Y}_{i\cdot}^{*}) = \frac{\sqrt{\widehat{Var}(\widehat{Y}_{i\cdot}^{*})}}{\widehat{Y}_{i\cdot}^{*}} \times 100\,,$$

$$RRMSE(\widehat{Y}_{i\cdot}^{*}) = \frac{\sqrt{mse(\widehat{Y}_{i\cdot}^{*})}}{\widehat{Y}_{i\cdot}^{*}} \times 100\,.$$

여기에서 만약 $\widehat{Y}_{i\cdot}^{*}$가 불편추정량이면, $RSE(\widehat{Y}_{i\cdot}^{*})$와 $RRMSE(\widehat{Y}_{i\cdot}^{*})$는 동일한 값을 갖는다.

충북지역의 경제활동인구조사 자료로부터 10개 행정자치구역들에 대한 실업자 총계 추정값들의 RB, RSE 와 RRMSE 값들을 계산하면 다음 표 5.2 와 같이 주어진다.

표 5.2 에서 합성추정값들과 복합추정값들의 편향 값들을 비교해 보면, 복합추정값들의 절대 편향값들의 평균 ($Av.RB$)이 10.26%로 합성추정값들의 절대 편향값들의 평균인 12.24% 보다 다소 작은 값을 나타낸다. Area 3과 Area 5를 제외한 나머지 8개 지역들에 대해서는 합성추정 및 복합추정값들의 상대 편향값들은 비교적 큰 값을 나타낸다. 따라서 합성추정 및 복합추정량의 신뢰성은 이들 추정량들의 편향과 추정오차의 영향이 함께 고려되어 평가되어야 한다.

우리는 실업자 총계 추정값들의 RRMSE(또는 RSE)값을 통해 소지역 추정값들의 신뢰성을 평가 하고자 한다. 한편, 표 5.2 에서 직접추정량은 근사적인 불편추정량이므로 직접추정값들의 RSE 값들은 RRMSE 값과 근사적으로 동일하다. 소지역들에 대한 RRMSE 값

표 5.2: 충북지역의 시군 단위 행정자치구역들에 대한 실업자 총계 추정값들의 RSE, RB, RRMSE 값 (2000년 12월)

Area No.	직접추정법	합성추정법		복합추정법	
	RSE_i	RB_i	$RRMSE_i$	RB_i	$RRMSE_i$
1	20.35	6.92	7.27	5.99	6.15
2	36.59	23.77	25.69	18.39	19.91
3	106.91	−2.95	5.99	−2.87	5.89
4	74.15	16.26	38.30	14.37	33.73
5	86.58	−7.04	14.51	−6.67	13.84
6	45.23	17.56	22.90	14.43	18.80
7	62.56	14.86	20.25	13.29	18.21
8	49.77	15.25	18.97	13.49	16.78
9	58.83	15.01	19.88	10.20	13.50
10	74.93	−2.75	10.79	−2.82	10.79
Av.RB		12.24		10.26	
Av.RSE	61.59				
Av.RRMSE			18.46		15.73

Av.RB= RB의 절대값들의 평균.
Av.RSE= RSE 값들의 평균.
Av.RRMSE= RRMSE 값들의 평균.

들의 허용한계 기준을 25%로 할때, 직접추정값들의 RSE(=RRMSE) 값들은 Area 1을 제외하고는 허용한계의 기준치를 만족하지 않는다. 반면, 합성추정 및 복합추정값들의 RRMSE 값들은 모든 소지역들에 대해서 직접추정값들의 RSE(=RRMSE) 값들보다 현저히 작은 값을 나타내며, 합성추정값들은 Area 2와 Area 5를 제외한 나머지 8개 지역들에서 RRMSE의 허용한계 기준치를 만족하고, 복합추정값들은 Area 5를 제외한 모든 지역들에서 RRMSE의 허용한계 기준치를 만족한다. 모든 소지역들에 대해서 합성추정 및 복합추정값들이 상당히 신뢰할 만한 추정결과를 나타내며, 효율이득 면에서는 복합추정값들이 합성추정값들에 비해 상대적으로 높은 효율이득을 나타낸다. 따라서 현 경제활동인구조사 체계 하에서는 복합추정량이 다른 추정량들에 비해 안정성과 신뢰성이 매우 높다는 사실을 충북의 시군 단위 소지역 추정결과에서 확인할 수 있다.

6. 결 론

우리나라의 경제활동인구조사는 전국단위의 대규모 통계조사로써 매월 취업자와 실업자 총계 등을 추정하는 유일한 공식 자료원이다. 매월 발표되는 자료는 실업률, 취업률, 경

제활동참가율 등과 같은 일에 종사할 수 있는 인구의 인구사회적인 특성들에 대한 정보도 포함된다. 그러나 경제활동인구조사는 대영역인 광역시 또는 도 단위들의 통계 작성을 목적으로 실시되기 때문에 최근들어 사회적 이슈가 되고 있는 시군구 단위의 행정자치구역들과 같은 부차관심영역들에 대한 실업통계는 경제활동인구조사 자료만으로는 추정이 불가능하다.

우리는 경제활동인구조사 자료와 통계청의 공식자료 만을 이용하여 대영역 내의 시군구 단위 행정자치구역들에 대한 실업자 총계를 추정하기 위해 설계기반 간접추정량으로써 합성추정량과 복합추정량을 제안하였다. 시군구 단위 소지역 추정값들의 정확도의 측도로써 추정값들의 잭나이프 평균제곱오차와 RRMSE가 제시되었다. 경제활동인구조사 자료를 이용한 추정결과에 의하면 시군구 단위 소지역 추정에 있어서 합성추정량과 복합추정량이 직접추정량에 비해 월등한 효율을 나타냈으며, 이러한 추정량들은 대부분 소지역 추정값들의 목표허용오차 한계를 만족하였다. 효율이득면에서는 복합추정량이 다른 추정량들에 비해 탁월하였다. 그러나 현재 통계청에서 실업자 총계 추정에 적용하고 있는 직접추정량은 현행 경제활동인구조사 체계 하에서 시군구 단위 소지역추정에 적용하기에는 무리가 있었다.

우리나라의 경제활동인구조사는 매월 주기적으로 실시되며 매 5년 단위로 표본 개편이 이루어진다. 새로운 표본설계에서는 시군구 단위 소지역 통계의 신뢰성이 확보될 수 있도록 모집단의 층화, 표본배정, 집락화의 수준 등에 대한 전반적인 검토가 요구된다. 또한 새로운 표본에서는 대영역 내의 시군구 단위 소지역들 뿐만 아니라 성별, 연령대별, 학력별 등과 같은 세부 관심영역들에 대한 시군구 단위 통계작성도 가능하도록 검토가 이루어져야 할 것이다.

참고문헌

[1] 이계오 (2000). 시군구 실업자 추정을 위한 소지역 추정법, <응용통계연구>, 제13권 2호, 275-286.

[2] Bender, R.K. (1985). Experience with small area population estimates, *Survey Methodology*, 11, 219-222.

[3] Decaudin, G., and Labat, J.C. (1997). A synthetic, robust and efficient method of making small area population estimates in France, *Survey Methodology*, 23, 91-98.

[4] 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.

[5] Falorsi, P.D., Falorsi, S., and Russo, A. (1994). Empirical comparison of small area estimation methods for the Italian Labour Force Survey, *Survey Methodology*, 20, 171-176.

[6] Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: An appraisal, *Statistical Science*, 9, 55-93.

[7] Hidiroglou, M.A., and Sarndal, C.E. (1985). An empirical study of some regression estimators for small domains, *Survey Methodology*, 11, 65-77.

[8] Marker, D.A. (1999). Organization of small area estimators using generalized linear regression framework, *Journal of Official Statistics*, 15, 1-24.

[9] Prasad, N.G.N., and Rao, J.N.K. (1990). The estimation of the mean squared error of small-area estimators, *Journal of the American Statistical Association*, 85, 163-171.

[10] Purcell, N.J., and Kish, L. (1979). Estimation for small domains, *Biometrics*, 35, 365-384.

[11] Singh, M.P., Gambino, J., and Mantel, H.J. (1994). Issues and strategies for small area data, *Survey Methodology*, 20, 3-22.

[12] Statistics Canada (1998). *Methodology of the Canadian Labour Force Survey*, Catalogue 71-526-XPB, Statistics Canada, Ottawa.

[13] Wolter, K.M. (1985). *Introduction to Variance Estimation*, Springer-Verlag.

[2002년 5월 접수, 2002년 12월 채택]

Design-Based Small Area Estimation for the Korean Economically Active Population Survey

Yeon Soo Chung [1] Kay-O Lee [2] Woo Il Lee [3]

ABSTRACT

In this study, we suggest the method of small area estimation based on the Economically Active Population Survey (EAPS) data in producing unemployment statistics for the local self-government areas (LSGAs) within large areas. The small area estimators considered are design-based indirect estimators such as the synthetic and composite estimators. The jackknife mean square error was used as a measure of accuracy of such small area estimators. The total unemployed and jackknife mean square errors of the 10 LSGAs within the large area of ChoongBuk region are derived from the estimation procedure suggested in this study, using EAPS data of December 2000. The reliability of small area estimators was assessed using the relative bias values and relative root mean square errors of these estimators. We find that under the current Korean EAPS system, the composite estimator turns out to be much more stable than other estimators.

Keywords: Small Area Estimation; Synthetic Estimator; Composite Estimator; Jackknife Mean Square Error; Relative Bias; Relative Mean Square Error

1) Associate Professor, Department of Computer Science and Statistics, Korea Air Force Academy.
E-mail:yschung@afa.ac.kr
2) Advisory Professor, Gallup Korea.
E-mail:kayolee@orgio.net
3) Assistant Professor, Department of Computer Science and Statistics, Korea Air Force Academy.
E-mail:wilee@afa.ac.kr