

시군구 실업자 추정에서 분산 추정

이계오¹⁾, 김규영²⁾

요 약

경제활동인구조사에서 시군구의 실업자를 추정하는데 소지역 추정법을 이용하는 방안에 대한 연구는 관심의 대상이 되고 있다. 본 연구에서는 합성 추정법과 복합 추정법을 이용한 시군구 실업통계 작성법을 소개하였고 추정량이 편향이므로 잭나이프 방법을 이용한 추정량의 정도를 계산하는 절차를 설명하였으며, 광주광역시의 구별 실업통계작성을 사례로 제시하였다.

주요용어 : 합성 추정량, 복합 추정량, MSE의 잭나이프 추정

1. 서론

경제활동(이하 경활)인구조사를 위해서 통계청에서 매월 3만여 표본가구를 대상으로 15일이 포함된 주간을 기준으로 경제활동에 관련된 사항을 조사하여 특·광역시와 도단위로 경활인구에 관한 통계를 25일경에 발표하고 있다. 그러나 지방자치의 기초단위인 시군구에 대한 실업통계가 요청되고 있으나 각 시군구별로 배정된 표본조사구 수가 적을 뿐 아니라 조사된 조사구 수의 분포가 불균형적이어서 신뢰도가 유사한 시군구별 실업통계의 생산이 현재의 통계적 기법으로는 어렵다. 이와같이 표본설계당시에는 통계생산단위로 고려하지 않았으나 통계조사를 실시한 후에 좀더 세분화된 단위에 대한 통계를 생산하는데 적용되고 있는 통계생산기법중 하나가 소지역 추정법이며 이중에서 시군구단위의 실업통계를 생산하는데 이용될 수 있는 합성 추정법과 복합 추정법은 주변 유사 시군구의 조사정보를 통합하여 통계적 특성이 유사하다는 가정을 전제로 하고 있기 때문에 이들의 가정조건이 완전하게 성립하지 않을때는 이들 추정법에 의해서 산정한 추정량은 편향 추정량이 되기 때문에 합성 추정량과 복합 추정량의 평균제곱오차를 계산하기 위해서 잭나이프방법을 이용하는 절차를 살펴보았다. 본 연구의 실용적 타당성을 알아보기 위해서 실제 경활조사 자료를 이용하여 광주광역시의 구별 실업통계를 생산하였으며 각 추정법을 수치적인 방법으로 비교하였다.

2. 소지역 추정법

소지역 추정법은 통계생산 단위에 대한 표본수가 적거나 아예 표본이 없는 경우에 주변의

1) 공군사관학교 전산통계학과 교수,(363-849) 충북 청원군 남일면 쌍수리사서함335-2호

2) 통계청 사무관,(302-701) 대전시 서구 둔산동 920

조사정보를 통합하고, 통계적 특성이 유사하다는 가정하에서 세분화된 통계생산단위의 통계를 생산하는 방법으로서 미국이나 캐나다 등에서 많이 이용되고 있는 통계 작성법이다. 본 연구에서는 직접 추정법, 합성 추정법과 복합 추정법을 다루겠다.

2.1. 직접 추정법(Direct Estimation)

직접추정량(direct estimator)은 보통 해당 소지역에서 조사된 자료만을 이용하여 추정되며, 간혹 센서스나 행정자료로부터 획득된 보조정보를 조사자료에 추가하여 추정하기도 한다. 가장 간단한 총계추정에 대한 직접추정량으로써 다음과 같은 단순추정량(expansion estimator)을 들 수 있다.

$$\hat{Y}_{e,a} = \sum_{i \in s_a} \omega_i y_i \quad (2.1)$$

여기에서 s_a 는 소지역 a 의 표본들의 집합, ω_i 는 조사단위 i 에 대한 가중치를 나타낸다. 위의 직접추정량은 불편추정량이나 소지역 a 의 표본크기가 작을 경우에는 분산이 커지기 때문에 신뢰성에 문제가 있을 수 있다.

소지역 a 의 모집단의 크기 N_a 를 알고있을 경우에는 다음과 같은 사후총화추정량이 이용될 수 있다.

$$\begin{aligned} \hat{Y}_{pst,a} &= N_a \frac{\sum_{i \in s_a} \omega_i y_i}{\sum_{i \in s_a} \omega_i} \\ &= N_a \bar{y}_{e,a} \end{aligned} \quad (2.2)$$

위의 사후총화추정량은 먼저 언급된 단순추정량보다는 안정적이다.

2.2 합성 추정법(Synthetic Estimation)

합성추정법(synthetic estimation)은 소지역을 포함하는 대영역의 정보를 함께 이용하는 방법으로써 소지역과 대영역의 특성 구조가 유사하다는 가정 하에서 이용된다. 합성추정량의 분산은 직접추정량의 분산에 비해 작으나 전체한 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.

$$\hat{Y}_{syn,m,a} = N_a \frac{\sum_{i \in s} \omega_i y_i}{\sum_{i \in s} \omega_i} = N_a \bar{y} \quad (2.3)$$

총화 또는 사후총화에 근거한 합성추정량은 보통 다음과 같은 형태로 주어진다.

$$\hat{Y}_{syn,st,m,a} = \sum_h N_{h,a} \frac{\sum_{i \in s_h} \omega_i y_i}{\sum_{i \in s_h} \omega_i} = \sum_h N_{h,a} \bar{y}_h \quad (2.4)$$

직접추정법에서와 마찬가지로 합성추정법에서도 비합성추정법(ration synthetic estimation)이 고려될 수 있다. 비합성추정법은 모집단의 크기 N_a 또는 $N_{h,a}$ 외에 소지역 추정을 위한 보조정보로써 공변량 x 를 이용하여 추정량(ratio synthetic estimator)은 다음과 같은 형태로 정의된다.

$$\hat{Y}_{syn,r,a} = X_a \frac{\hat{Y}_e}{\hat{X}_e}, \quad \hat{Y}_{syn,st,r,a} = \sum_h X_{h,a} \frac{\hat{Y}_{h,e}}{\hat{X}_{h,e}}, \quad (2.5)$$

여기에서 $\hat{Y}_e = \sum_{i \in S} \omega_i y_i$ 는 y 에 대한 모집단 총계 추정량, $\hat{Y}_{h,e} = \sum_{i \in S_h} \omega_i y_i$ 는 층의 총계 추정치를 나타낸다.

2.3 복합 추정법(Composite Estimation)

복합추정량(composite estimator)은 직접추정량의 불안정성과 합성추정량의 잠재적 편향 가능성을 보완하기 위해 두 추정량의 가중평균을 취하며 일반적인 형태는 다음과 같이 주어진다.

$$\hat{Y}_{com,a} = \lambda_a \hat{Y}_{dir,a} + (1 - \lambda_a) \hat{Y}_{syn,a}, \quad (2.6)$$

여기에서 가중치 λ_a 는 적절히 선택되는 값이다.

가중치 λ_a 는 결정하는 방법은 크게 세가지 정도로 구분될 수 있다. 첫 번째 방법은 가장 간단한 방법으로써 가중치 λ_a 를 고정계수로 두는 방법인데 추정량의 신뢰성에 문제가 있어 많이 사용되지는 않는다. 두 번째 방법은 추정하고자하는 소지역의 표본크기를 반영하는 방법이다. 가중치를 결정하는 세 번째 방법은 직접추정량과 합성추정량의 평균제곱오차와 두 추정량의 공분산을 자료로부터 추정하여 최적가중치를 산정하는 방법이다. 복합추정량의 평균제곱오차는 다음 식과 같이 나타낼 수 있다.

$$\begin{aligned} MSE(\hat{Y}_{com,a}) &= \lambda_a^2 MSE(\hat{Y}_{dir,a}) + (1 - \lambda_a)^2 MSE(\hat{Y}_{syn,a}) \\ &\quad + 2\lambda_a(1 - \lambda_a)E(\hat{Y}_{dir,a} - Y_a)(\hat{Y}_{syn,a} - Y_a) \end{aligned} \quad (2.7)$$

$\hat{Y}_{com,a}$ 의 MSE 를 최소화하는 가중치 λ_a 다음 식과 같이 주어질 수 있다.

$$\lambda_a = \frac{MSE(\hat{Y}_{syn,a}) - E(\hat{Y}_{syn,a} - Y_a)(\hat{Y}_{dir,a} - Y_a)}{MSE(\hat{Y}_{syn,a}) + MSE(\hat{Y}_{dir,a}) - 2E(\hat{Y}_{syn,a} - Y_a)(\hat{Y}_{dir,a} - Y_a)} \quad (2.8)$$

3. 평균제곱오차의 잭나이프 추정

잭나이프 추정법에 의한 복합 추정량의 MSE 추정에 대한 구체적인 절차는 아래와 같다.

(절차 1) 각 구별로 모든 표본자료를 이용하여 추정법에 따른 추정식 을 이용하여 실업자를

시군구 실업자 추정에서 분산 추정

추계하고, 각 구에서 조사구를 하나씩 제외한 후에 해당식을 각각 이용하여 구별 실업자를 추정한다.

(절차 2) 각 구별로 합성 추정량의 MSE 잭나이프 추정값을 다음식으로 계산한다.

$$\begin{aligned} \text{mse}(\hat{Y}_{i \cdot}^S) &= \{ \widehat{Bias}_{Jack,i}(\hat{Y}_{i \cdot}^S) \}^2 + \widehat{Var}_{Jack,i}(\hat{Y}_{i \cdot}^S) \\ &= (n_i - 1)^2 \left(\frac{\sum_{k=1}^{n_i} \hat{Y}_{i \cdot}^S(k)}{n_i} - \hat{Y}_{i \cdot}^S \right)^2 \\ &\quad + \frac{n_i - 1}{n_i} \sum_{k=1}^{n_i} \left(\hat{Y}_{i \cdot}^S(k) - \frac{\sum_{k=1}^{n_i} \hat{Y}_{i \cdot}^S(k)}{n_i} \right)^2 \end{aligned}$$

(절차 3) (절차 2)에서 계산한 결과를 이용하여 복합 추정량에 대입할 가중값을 계산한다.

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

(절차 4) 각 구별로 복합 추정량의 MSE잭나이프 추정값을 아래의 식으로 산정한다.

$$\begin{aligned} \text{mse}(\hat{Y}_{i \cdot}^C) &= \{ \widehat{Bias}_{Jack,i}(\hat{Y}_{i \cdot}^C) \}^2 + \widehat{Var}_{Jack,i}(\hat{Y}_{i \cdot}^C) \\ &= (n_i - 1)^2 \left(\frac{\sum_{k=1}^{n_i} \hat{Y}_{i \cdot}^C(k)}{n_i} - \hat{Y}_{i \cdot}^C \right)^2 \\ &\quad + \frac{n_i - 1}{n_i} \sum_{k=1}^{n_i} \left(\hat{Y}_{i \cdot}^C(k) - \frac{\sum_{k=1}^{n_i} \hat{Y}_{i \cdot}^C(k)}{n_i} \right)^2 \end{aligned}$$

(절차 5) (절차 3)과 (절차 4)에 의해서 계산된 MSE 추정값을 이용하여 구별 실업자의 추정값을 보정식에 대입하여 최종적으로 실업자를 추계한다.

4. 광주광역시의 구별 실업 통계 작성사례

구별 실업자 추정량은 직접 추정량, 합성 추정량과 복합 추정량에 대해서만 살펴보자
가. 직접 추정량

$$\begin{aligned} \hat{Y}_{i \cdot} &= \sum_{s=1}^2 {}_s\hat{Y}_{i \cdot} \quad , \quad i=1,2,\dots,I; s=1,2; h=1,2,\dots,n_i \\ &= \sum_{s=1}^2 \sum_{h=1}^{n_i} {}_sM_i {}_sY_{ih} \end{aligned} \quad (4.1)$$

여기에서 s 는 성별(남-여)이며, n_i 는 경제활동인구조사에서 i 번째 지역의 표본조사구 수, ${}_sY_{ih}$ 는 각 성별에서 i 번째 지역의 표본 조사구에서 조사한 실업자 수를 나타낸다.

직접추정량 $\hat{Y}_{i.}$ 의 분산은 다음 식과 같이 주어진다.

$$Var(\hat{Y}_{i.}) = \sum_{s=1}^2 Var({}_s \hat{Y}_{i.}) + 2Cov({}_1 \hat{Y}_{i.}, {}_2 \hat{Y}_{i.}), \quad i=1,2,\dots,I$$

나. 합성 추정량

대영역을 I 개의 시군구 단위의 소지역들로 분할하고, 대영역을 특성 기준에 따라 유사성을 갖는 J 개의 성별-연령대별 범주들로 구분할 때 i 번째 소지역의 합성추정량 $\hat{Y}_{i.}^S$ 는 다음과 같이 주어질 수 있다.

$$\hat{Y}_{i.}^S = \sum_{j=1}^J \eta_{ij} \Psi^{a.j.}, \quad i=1,2,\dots,I, \quad (4.2)$$

식(4.2)에서 가중치 $\eta_{ij} = (\xi_{ijt}^C / \xi_{ijt}^R) \xi_{ij} s x_j$ 는 i 번째 소지역에서 j 번째 범주에 대한 경제활동 추정인구를 나타낸다.

소지역 i 에서 j 번째 범주에 대한 경제활동인구 η_{ij} 를 상수로 가정한다면 합성추정량 $\hat{Y}_{i.}^S$ 의 분산은 다음 식과 같이 주어질 수 있다.

$$Var(\hat{Y}_{i.}^S) = \sum_{j=1}^J \eta_{ij}^2 Var(\Psi^{a.j.}) + 2 \sum_{j_1 < j_2} \eta_{ij_1} \eta_{ij_2} Cov(\Psi^{a.j_1.}, \Psi^{a.j_2.}) \quad i=1,2,\dots,I \quad (4.3)$$

다. 복합 추정량

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

$$\hat{Y}_{i.}^C = \hat{\omega}_{i(opt)} \hat{Y}_{i.} + (1 - \hat{\omega}_{i(opt)}) \hat{Y}_{i.}^S, \quad i=1,2,\dots,I \quad (4.4)$$

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

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

위에서 언급한 직접 추정량, 합성추정량과 복합 추정량에 의해서 조사된 자료를 근거로 계산된 추정값을 기준으로 하여 통계청에서 발표하는 광주광역시의 2001년 5월 실업자수와 일치시키기 위해서 다음과 같은 비추정식을 이용하여 보정하였다.

$$\hat{Y}_{i.}^A = \left(\frac{\hat{Y}_{i.}^*}{\sum_i \hat{Y}_{i.}^*} \right) \hat{Y} \quad (4.6)$$

식 (4.6)에 의해서 직접 추정값, 합성 추정값과 복합 추정값의 계산결과는 <표1>에 주어졌다.

<표 1 > 광주광역시의 구별 추정값과 상대표준오차(2001년 5월)

시군구	직접추정법			합성추정법				복합추정법			
	$\hat{Y}_{i.}$	S.E	RSE	$\hat{Y}_{i.}$	S.E	MSE	RSE	$\hat{Y}_{i.}$	S.E	MSE	RSE
동구	2,807	976	34.77	2,285	281	78,961	10.65	2,397	270	79,324	10.18
서구	4,837	1,164	24.06	5,168	627	393,129	10.50	5,165	552	402,243	9.66
남구	5,289	1,137	21.50	4,253	518	268,324	10.54	4,499	471	321,876	9.46
북구	8,629	1,256	14.56	8,602	1,046	1,094,116	10.53	8,498	804	947,733	8.55
광산구	2,761	830	30.06	4,015	482	232,324	10.39	3,764	417	254,654	10.01
합계	24,323			24,323				24,323			

5. 결론

소지역 추정법을 이용하여 시군구단위의 실업자를 추정하는데 표본조사구를 획기적으로 증대시키지 않더라도 어느 정도의 신뢰수준을 갖춘 통계를 생산할 수 있음을 보였고 특히 합성추정법이나 복합추정법에서 잭나이프 추정법으로 MSE를 추정할 수 있음을 수치적으로 보였다. 전국적으로 시군구를 유사층으로 층화하여 층내의 실업특성을 동일한 수준으로 만들 수 있다면 시군구단위의 경찰통계는 표본조사구를 약간만 증대한다면 전국 234개 시군구에 대해서 정책입안의 참고정보로 활용할 수준으로 생산 가능할 것이다. 그러나 본 연구는 가능성을 보여주는 의미가 있을뿐 이의 실용화를 위해서 프로그램개발등 심층적인 연구가 필요할 것이다

참 고 문 헌

- [1] 이계오의 2인(1999), 「표본조사론」 (한국방송대학교 출판부)
- [2] 통계청(1997), “가구부문 표본개편결과 보고서”
- [3] 통계청(2000), “소지역 추정법 연구”, 통계기획국 조사관리과
- [4] 통계청(2001), “캐나다 노동력조사의 방법론”, 통계기획국 조사관리과
- [5] 한국노동연구원(1999), “실업통계의 개선방안”, 강순희, 전재식, 이계오
- [6] Ghosh,M. and Rao,J.N.K. (1994) Small area estimation : an appraisal. Statistical Science, 9, 55 - 93.
- [7] Lahiri, P., and Rao,J.N.K. (1995). Robust estimator of mean square error of small area estimators, Journal of American Statistical Association, 90, 758-766