The Korean Communications
in Statistics Vol. 11 No. 3, 2004
pp. 583-595

Small Area Estimation Techniques Based on Logistic Model to Estimate Unemployment Rate[1)]

Young-Won Kim[2)] and Hyung-a Choi[3)]

Abstract

For the Korean Economically Active Population Survey(EAPS), we consider the composite estimator based on logistic regression model to estimate the unemployment rate for small areas(Si/Gun). Also, small area estimation technique based on hierarchical generalized linear model is proposed to include the random effect which reflect the characteristic of the small areas. The proposed estimation techniques are applied to real domestic data which is from the Korean EAPS of Choongbuk. The MSE of these estimators are estimated by Jackknife method, and the efficiencies of small area estimators are evaluated by the RRMSE. As a result, the composite estimator based on logistic model is much more efficient than others and it turns out that the composite estimator can produce the reliable estimates under the current EAPS system.

Keywords : composite estimator, hierarchical generalized linear model, logistic regression. unemployment rate,

1. 서론

경제활동인구조사를 비롯한 대부분의 정부통계를 생산하기 위한 표본설계는 광역시 또는 도 단위와 같은 대영역의 통계를 생산할 목적으로 설계되기 때문에, 시군구 등과 같은 소지역의 경우 배정되는 표본 조사구수가 극히 적어 신뢰할 수 있는 통계 산출이 어렵다. 하지만 비용을 고려할 때 전국의 모든 시군구 통계를 고려한 새로운 표본조사를 실시하는 것은 현실적으로 불가능하다. 따라서 기존 표본설계에서 조사된 자료를 가지고 일정 수준의 정도(precision)를 만족하는 시군구 단위의 통계를 생산할 수 있는 소지역 추정기법에 대한 연구가 요구된다. 소지역 추정기법이란 배정된 표본크기가 작은 소지역(small area)이나 성별, 연령, 교육수준, 소득수준 등과 같은 변수의 특성으로 분류된 소영역(small domain)에 대한 통계를 생산하는데 이용되는 추정방법이다.

우리나라에서도 지방지차제의 실시 등에 따라 다양한 행정부처에서 지역별 정부정책 수립을 위

1) This Research was supported by the Sookmyung Women's University Research Grants 2003.
2) Professor, Department of Statistics, Sookmyung Women's University, Seoul, 140-742, Korea. E-mail: ywkim@sookmyung.ac.kr
3) Graduate Student, Department of Statistics, Sookmyung Women's University, Seoul, 140-742, Korea.

해 시군구 등 소지역 실업률에 대한 정확한 통계를 필요로 하고 있다. 그러나 우리나라에서 실업률 산출을 목적으로 실시되는 경제활동인구조사 표본설계는 특광역시 및 도별 실업자 통계 산출을 목적으로 하기 때문에 시군구와 같은 소지역에 대한 통계 산출은 표본설계에 반영되어 있지 않다. 따라서 경제활동인구조사 자료를 이용하여 시군구 실업률 통계를 직접 생산할 경우 각 시군구별 표본 조사구수가 너무 작기 때문에 신뢰할만한 통계 산출이 불가능하다.

외국에서는 분석대상에 따라 다양한 형태의 소지역 추정기법들에 대한 심층적인 연구들이 진행되고 있다. 특히 미국, 캐나다, 영국 등의 경우 정부기관과 전문 학자들과의 공동연구를 통해 소지역 추정기법에 대한 연구가 활발히 추진되고 있다. Rao(2003)은 다양한 형태의 소지역 추정방법에 대한 최근까지의 연구결과들을 체계적으로 정리하고 있다. 우리나라의 경우 김영원과 성나영(2000)이 도소매업 사업체 조사를 바탕으로 소지역 추정기법의 도입 가능성을 검토한 이후 박종태와 이상은(2001)은 경제활동인구조사를 토대로 경기도 시군구의 실업자 총계 추정문제를 제한적인 모의실험을 통해 다루었다. 특히 최근에 들어 우리나라 정부기관에서도 소지역 추정기법 도입의 필요성을 절감하고, 관련 학자들과 연계하여 연구를 추진하고 있다. 현재까지 진행된 대표적인 연구로는 이계오, 류제복, 김영원(2001)이 수행한 소지역 실업자 총계 추정 가능성 검토를 위한 기초연구가 있으며, 정연수, 이계오, 이우일(2003) 및 Chung, Lee, Kim(2003)이 통계청의 협조를 받아 수행한 충청북도 시군구 실업자 총계 추정에 대한 연구가 있다. 정연수 등(2003) 및 Chung 등(2003)은 시군구 실업률 공식 통계 산출을 염두에 두고 통계청에서 본격적으로 추진한 연구라는 점에서 의의를 찾을 수 있다.

현재까지 우리나라에서 실업통계 관련 소지역 추정 연구에서는 실업자 총계 추정에 대한 연구는 활발하나 또 다른 관심대상이 될 수 있는 실업률에 대한 추정문제에 대해서는 연구가 미흡하다. 우리나라 실업통계 소지역 추정 문제와 관련된 대표적인 연구인 Chung 등(2003)에서는 결과적으로 시군구에 대한 실업자 총계와 경제활동인구 총계를 별도의 소지역 추정방법을 적용하여 산출하고, 이들의 비(ratio)를 구하여 소지역 실업률을 산출할 수는 있지만 실업률 추정 자체를 목적으로 하는 경우 이는 효율적인 방법이라 볼 수 없다. 이런 접근방식은 우리나라 관련 분야 연구자들이 실업통계와 관련된 소지역 통계기법이 활발히 연구되고 있는 캐나다의 실업률 소지역추정기법(Drew 등(1982), Rao(2003) 참조)에 많은 영향을 받았기 때문이라고 판단된다.

하지만 지역별 실업자 수가 아니라 실업률 자체가 주 관심대상인 경우 좀더 효율적이고 안정적인 소지역 실업률 산출을 위해서는 소지역별 실업률을 직접 활용하는 소지역 추정기법이 대안이 될 수 있다. 이런 관점에서 김영원과 조란(2004)은 경제활동인구 자료를 바탕으로 실업자 총계 대신 실업률을 추정할 수 있는 합성추정법을 제시하고, 이를 통해 보다 안정적인 실업통계 산출이 가능하다는 연구결과를 제시하고 있다. 한편 실업률을 연구대상으로 하는 경우 이항자료에 활용할 수 있는 소지역 추정모형의 활용이 필요하다. 이와 관련하여 Ghosh 등(1998)은 혼합 일반화선형모형(mixed generalized linear model)을 토대로 한 계층적베이즈(hierarchical Bayes) 추정법을 제시하였고, Jiang 등(2001)은 분산안정화 변수변환을 적용하여 Fay-Herriot 모형을 활용하는 방안을 연구하였다. 특히 Ambler(2001)를 참고로 하면 영국 통계국(ONS)에서는 지역별 실업률 추정을 위해 이항분포를 기반으로 한 소지역 추정기법을 연구 개발하고 있다는 것을 확인할 수 있다.

본 연구에서는 2000년 12월 경제활동인구조사 자료를 바탕으로 한 충청북도내의 10개 시군에 대한 실업률 추정 문제를 다룬다. 보다 효율적인 소지역 실업률 추정을 위해 본 연구에서는 우선 로지스틱모형을 기초로 한 합성추정량을 제시하고, 이를 이용한 복합추정방법을 제시한다. 여기서 사용된 로지스틱 모형을 기초로 한 소지역 추정량은 대지역에서 성별-연령별 실업률을 산출하여

활용하는 Chung 등(2003)과 김영원 등(2004)이 제시한 기존의 실업통계 소지역 추정법과는 달리, 이항(binomial) 자료에 적합한 모형기반(model based) 합성추정량을 활용한 새로운 간접추정법이라고 볼 수 있다.

아울러 본 연구에서는 지역 랜덤효과(random effect)를 포함한 GLMM(generalized linear mixed model)을 활용하기 위해, Ghosh 등(1998)이 제시한 HB(hierarchical Bayes) 추정법을 효과적으로 대체할 수 있을 것으로 판단되는 Lee와 Nelder(1996)가 제시한 HGLM(hierarchical GLM)을 시군구 실업률 추정에 적용하는 방안도 함께 제시한다. 제시된 추정법의 효율성 검증을 위해 충청북도 10개 시군구 단위 소지역에 대해 소지역 실업률을 추정하고, 잭나이프 방법을 사용하여 평균제곱오차(MSE)를 구해 각 추정방법의 효율성을 비교 분석한다.

2. 경활 자료를 이용한 소지역 실업률 추정

2.1 직접추정량

소지역 i에 대한 실업률을 추정하기 위한 직접추정량은 해당 소지역에 배정된 표본조사구에서 얻어진 자료만을 이용한다. 현행 통계청의 경제활동인구조사 체계에서의 소지역 i의 실업률에 대한 직접추정량 $\widehat{p_{i.}^{d}}$은 다음과 같다.

$$\widehat{p_{i.}^{d}} = \frac{\sum_{s=1}^{2}\sum_{j=1}^{n_i} {}_s\widehat{Y}_{ij}}{\sum_{s=1}^{2}\sum_{j=1}^{n_i} {}_s\widehat{X}_{ij}} = \frac{\sum_{s=1}^{2}\sum_{j=1}^{n_i} {}_sM_i\,{}_sY_{ij}}{\sum_{s=1}^{2}\sum_{j=1}^{n_i} {}_sM_i\,{}_sX_{ij}}, \quad i=1,...,I \tag{2.1}$$

여기서 s는 성별(남, 여)를 나타내고, n_i는 경제활동인구조사에서의 소지역 i에 할당된 표본조사구 수, ${}_sY_{ij}$는 경제활동인구조사에서 소지역 i의 j번째 표본조사구에서 성별 15세 이상의 실업자 수, ${}_sX_{ij}$는 경제활동인구조사에서 소지역 i의 j번째 표본조사구에서 성별 15세 이상의 경제활동인구를 나타내며, 승수 ${}_sM_i$는 상주인구 자료를 이용해 산출된 경제활동인구조사 가중치이다.

2.2 로지스틱 합성추정량

소지역 실업률에 대한 합성추정량(synthetic estimator)으로는 다양한 간접추정량의 사용이 가능하다. 김영원 등(2004)의 경우 충청북도의 시군을 시와 군지역으로 구분하여 구성한 2개의 대지역에서 각각 성별-연령별 구분에 따른 실업률을 산출하고, 여기에 각 소지역의 성별-연령별 상주인구 구성비를 반영한 합성추정량을 사용하고 있다. 김영원 등(2004)이 제시한 합성 추정량의 대안으로 로지스틱 회귀모형을 기반으로 한 합성추정량을 고려할 수 있다.

본 연구에서는 Ambler 등(2001)이 제시한 모형과 유사한 형태의 로지스틱 모형에 근거한 합성추정량을 사용한다. 여기서는 실업률과 관련이 많은 것으로 판단되는 성별-연령별 범주를 고려한 로지스틱모형을 고려한다. 이에 따라 경제활동인구 자료에서 조사대상자를 성별과 나이(35세 이상

과 미만)로 구분하여 4개의 범주를 구성하고, 각 소지역에서 범주별(k = 1, 2, 3, 4) 실업률을 식 (2.1)과 같은 형식으로 산출하여 이를 바탕으로 다음과 같은 로지스틱 모형을 적용한다.

$$\log\left(\frac{p_{ik}}{1-p_{ik}}\right) = \chi_{ik}'\underline{\beta} + Z_{ik}\gamma \qquad (2.2)$$

여기서 p_{ik}는 i 소지역 k 범주의 실업률, x_{ik} 는 나이 및 성별을 나타내는 범주형 변량, Z_{ik}는 해당 범주 상주인구 비율을 나타낸다. 모형 (2.2)에 상주인구 비율을 나타내는 변량 Z_{ik}를 포함시키고 있다는 점에 유의할 필요가 있다. 실제 경활 자료 분석에서 이 변수를 포함한 완전모형과 이를 제외한 축소모형에 대한 비교를 통해 이 변수가 포함되는 것이 모형의 적합도를 향상시키는 것으로 판단되어 모형에 포함한 것이다. i소지역 k범주에서 표본추출 가중치 M_{ik}를 반영한 실업자수를 Y_{ik}^{*}(즉, $Y_{ik}^{*} = M_{ik}Y_{ik}$), 표본추출 가중치 M_{ik}를 반영한 경제활동 인구수를 n_{ik} 라고 하면, 이항변수로 볼 수 있는 Y_{ik}^{*}를 기초로 (2.2)의 로지스틱모형을 적합시켜 얻은 결과를 요약하면 <표 2.1>과 같다. 여기서 col4는 성별-연령별 범주형 변수 x_{ik}, col5는 해당 범주 상주인구 비율 Z_{ik}를 나타내며, 분석결과에서 x_{ik}와 Z_{ik}는 모두 유의하다는 것을 볼 수 있다. <표 2.1>의 결과에 따른 적합된 로지스틱 회귀모형을 기반으로 각 소지역에서 성별-연령에 따른 범주별 실업률을 예측(prediction) 형식으로 산출할 수 있다.

<표 2.1> 로지스틱 모형의 적합 결과

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Error	Standard Chi-Square	Pr > ChiSq
Intercept		1	-2.3631	0.0528	2000.2303	<.0001
col4	1	1	0.5121	0.0140	1339.4901	<.0001
col4	2	1	-0.3598	0.0138	676.6624	<.0001
col4	3	1	0.1975	0.0166	140.9872	<.0001
col5		1	-3.9103	0.2093	349.0512	<.0001

한편 소지역 i에서 k범주의 상주인구 비율을 나타내는 Z_{ik}와 실업률 p_{ik}의 관련성을 검토해 보기 위해 상관분석을 해 본 결과, 상관계수가 -0.33257로 실업률과 상주인구 비율 간에 매우 유의적인 음의 상관관계가 있음을 알 수 있다. 이는 특정 연령-성별 범주에 대한 인구 구성비율이 높은 소지역에서는 해당 범주의 실업률이 낮다는 것을 의미한다. 다시 말하면 어떤 소지역에서 특정 성별-연령별 범주에 대한 인구 구성비가 상대적으로 높다는 것은 이 소지역에 해당 범주에 속하는 인구에 대한 취업 기회가 상대적으로 양호하며, 이에 따라 이 지역으로 해당 범주 인구가 많이 유입되어 거주하고 있다는 것으로 해석할 수 있다. 따라서 이제까지 알려져 있지 않지만 성별-연령별 인구구성비는 실업률 산출을 위한 소지역 추정 모형 개발에 매우 효과적으로 활용될 수 있는 보조 변수라고 볼 수 있다. 향후 경제활동인구 분석과 관련해서 이에 대한 보다 심층적인 연구가 필요하다고 판단된다.

2.3 복합추정량

직접추정량의 경우 일부 소지역에서는 표본 조사구수가 매우 적기 때문에 표본오차가 크게 되고, 반면에 합성추정량의 경우 일반적으로 편향이 발생하게 된다. 따라서 이들을 서로 보완하기 위해 흔히 복합추정량(composite estimator)이 사용된다. 따라서 본 연구에서는 소지역 i에 대한 다음과 같은 복합추정량 $\widehat{p_i^c}$ 을 고려한다.

$$\widehat{p_{i.}}^c = \widehat{\omega_i}\widehat{p_{i.}}^d + (1-\widehat{\omega_i})\widehat{p_{i.}}^s \tag{2.3}$$

여기서 $\widehat{p_{i.}^d}$는 식(2.1)에 의한 직접추정량이고, $\widehat{p_{i.}^s}$ 는 식(2.2)에 의한 결과를 기초로 한 로지스틱 합성추정량이다. 한편 $\widehat{\omega_i}$는 0과 1사이의 값을 갖는 가중치로 복합추정량의 MSE를 최소로 하는 다음과 같은 최적 가중값을 사용한다(Ghosh and Rao, 1994).

$$\widehat{\omega_i}(opt) = \frac{mse_J(\widehat{p_{i.}}^s)}{mse_J(\widehat{p_{i.}}^d) + mse_J(\widehat{p_{i.}}^s)}, \qquad i = 1, \ldots, 10 \tag{2.4}$$

여기서 mse_J은 잭나이프 MSE 추정량을 나타낸다.

한편, 비추정량에 해당하는 직접추정량 $\widehat{p_{i.}^d}$ 와 합성추정량 $\widehat{p_{i.}^s}$에 대한 MSE는 다음과 같은 잭나이프 공식을 적용하여 추정한다. 소지역 추정에서 MSE의 잭나이프 추정문제에 대해서는 어떤 방법이 가장 효율적인지 아직 확실하게 규명되지 않았지만, 여기서는 Chung 등(2003)이 제시한 방법을 따르고자 한다. 합성추정량의 잭나이프 MSE 추정은 다음과 같이 정리될 수 있다.

$$mse_J(\widehat{p_{i.}}^s) = \widehat{Var}_J(\widehat{p_{i.}}^s) + [\widehat{Bias}_J(\widehat{p_{i.}}^s)]^2 \tag{2.5}$$

여기서

$$\widehat{Var}_J(\widehat{p_{i.}}^s) = \frac{n_i - 1}{n_i}\sum_{k=1}^{n_i}(\widehat{p_{i.}}^s(k) - \widehat{p_{i.}}^s)^2,$$

$$\widehat{Bias}_J(\widehat{p_{i.}}^s) = (n_i - 1)\left(\frac{1}{n_i}\sum_{k=1}^{n_i}\widehat{p_{i.}}^s(k) - \widehat{p_{i.}}^s\right)$$

이고, n_i는 소지역 i의 조사구수, $\widehat{p_{i.}^s}(k)$는 k번째 조사구를 제외하고 구한 소지역별 합성추정치를 나타낸다. 직접추정량에 대한 잭나이프 MSE 추정량도 유사한 방법으로 산출한다. 아울러 식(2.3)에 제시된 복합추정량의 효율성 검토를 위한 MSE도 동일한 형식의 잭나이프 방법을 적용한다. 복합추정량의 MSE 추정에 있어서 복합추정량에 적용되는 가중치 $\widehat{w_i}$는 잭나이프 수행과정에서 각 조사구를 삭제할 때마다 다시 구해지는 것이 필요하다. 하지만 본 연구에서는 이를 상수로 처리하여 이런 변동을 반영하지 못하고 있기 때문에 산출된 MSE는 실제 보다 과소 추정될 수 있다는 한계를 갖고 있다는 점에 유의하기 바란다.

2.4 계층적 일반화 선형모형(HGLM)

2.2절에 제시된 로지스틱 모형은 성별-연령별 범주 변수와 해당 범주 상주인구 비율을 반영하고

있지만 각 소지역별 특성은 반영하지 못한다는 한계를 갖고 있다. 이에 따라 소지역별 특성을 반영하기 위해 직접추정량과 로지스틱 합성추정량의 가중평균 형태로 표현되는 식(2.3)의 복합추정량을 고려하는 것이 필요하다.

한편 로지스틱 모형에서 소지역별 특성을 반영하는 다른 방법으로는 각 소지역의 특성을 설명하는 랜덤효과(random effect)를 식(2.2)의 로지스틱 모형에 직접 포함하는 방법을 고려할 수 있다. 이와 같이 랜덤 소지역 효과를 반영한 모형은 GLM에 랜덤성분이 추가된 GLMM(generalized linear mixed model)에 해당한다. 최근의 GLM을 활용한 소지역 추정 관련 연구를 살펴보면, 랜덤효과가 포함된 GLMM에서 모형 적합을 위해 Ghosh(1998)가 제시한 HB(hierarchical Bayes) 추정법을 이용하는 것이 일반적이다. Ghosh(1988)가 제시한 HB 추정법은 모형의 모수에 사전 분포를 가정하고, 다차원의 적분을 수행하기 위해 깁스 표본추출(Gibbs Sampling)과 같은 복잡한 계산과정을 거쳐야 한다는 적용상의 불편함을 갖고 있다. 따라서 본 연구에서는 이런 베이즈 접근법 대신에 고전적인 우도함수를 기초로 한 Lee와 Nelder(1996)가 제시한 HGLM(hierarchical GLM)을 이용하는 새로운 방식의 시군구 소지역 실업률 추정법의 활용 가능성을 제안하고자 한다.

로짓(logit) 연결함수(link function)를 사용한 GLM인 모형 (2.2)에 지역 특성을 나타내는 랜덤효과 u_i를 반영함으로써 로지스틱 모형에서 복합추정량의 특성을 갖는 소지역 추정결과를 얻을 수 있다. 이는 Ghosh와 Rao(1994)에서와 같이 일반적인 혼합선형모형(mixed linear model)을 적용하여 복합추정량에 해당하는 EBLUP(empirical best linear unbiased predictor)을 직접 도출하는 대표적인 소지역 추정기법 유도방식과 유사한 접근법을 GLMM에서 적용한 것이다. 한편 Lee와 Nelder(2001)는 랜덤효과를 포함한 GLM에서 기존의 우도함수를 변형한 h-liklihood를 활용한 HGLM이란 새로운 통계적 추론 기법을 제시하고 있으며, 이들의 추론방법은 통계 처리 소프트웨어인 GenStat에 의해 손쉽게 수행될 수 있다.

본 연구에서는 구체적으로 다음과 같은 HGLM을 활용한 실업률 추정기법을 적용한다. 표본추출 가중치 M_{ik}를 반영한 실업자수를 Y_{ik}^*, 경제활동인구수를 n_{ik}라고 하면, 이항변수로 볼 수 있는 Y_{ik}^*를 이용한 소지역 모형은 다음과 같이 랜덤효과를 포함한 GLMM으로 표현 될 수 있다.

$$f(y_{ik}) = exp(y_{ik} ln(\frac{p_{ik}}{1-p_{ik}}) + n_{ik} ln(1-p_{ik}) + ln(\binom{n_{ik}}{y_{ik}})) \tag{2.6}$$

$$\theta_{ik} = ln(\frac{p_{ik}}{1-p_{ik}}) = x_{ik}^T \beta + u_i$$

여기서 p_{ik}는 i 소지역 k 범주의 실업률, x_{ik}는 소지역 추정을 위해 보조정보를 포함한 공변량, β는 성별, 연령, 성별-연령 교차 효과 및 k범주의 상주인구 비율을 나타내는 모수 벡터이다. u_i는 i 소지역의 특성을 나타내는 랜덤효과를 나타낸다. 여기서 u_i는 $N(0, \sigma_u^2)$으로 가정한다.

실업률 산출을 위한 모형에서는 공변량 x_{ik}는 성별 및 연령 효과를 나타내는 동시에 i 소지역의 k 범주의 상주인구 구성비를 포함한 것으로 식(2.6)에서 θ_{ik}는 구체적으로 다음과 같다.

$$\theta_{ik} = \mu + \tau_a + \tau_s + \tau_{as} + Z_{ik}\gamma + u_i$$

여기서, μ는 일반효과, τ_a는 연령효과, τ_s는 성별효과, τ_{as}는 성-연령 교차효과를 나타내고, Z_{ik}는 i지역 k범주 상주인구비율을 나타낸다. GenStat을 사용한 분석결과는 <표 2.2>와 같다.

<표 2.2> HGLM 적합 결과

		estimate	s.e.	t
1	Constant	-1.7790	0.1747	-10.18
2	age 2	-0.8855	0.0231	-38.27
3	sex 2	-0.3291	0.0206	-15.98
4	age 2 .sex 2	0.3392	0.0310	10.95
5	x	-4.1618	0.2037	-20.43

Note: s.e.s assume dispersion = 1.000

		estimate	s.e.	t
1	lambda 1	-1.2576	0.4724	-2.662

Note: s.e.s assume dispersion = 2.000

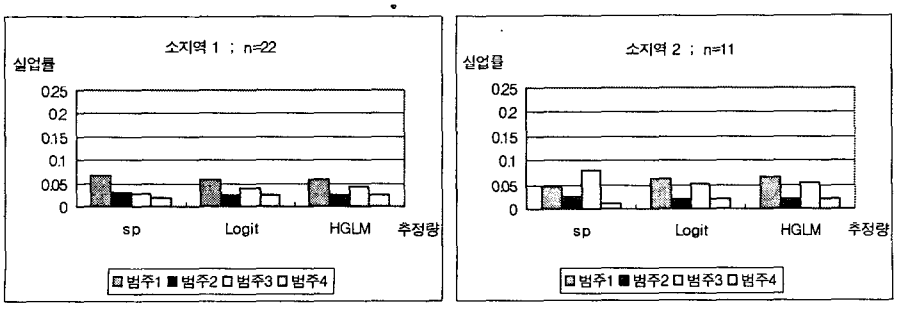
3. 성별-연령별 소지역 실업률 분석

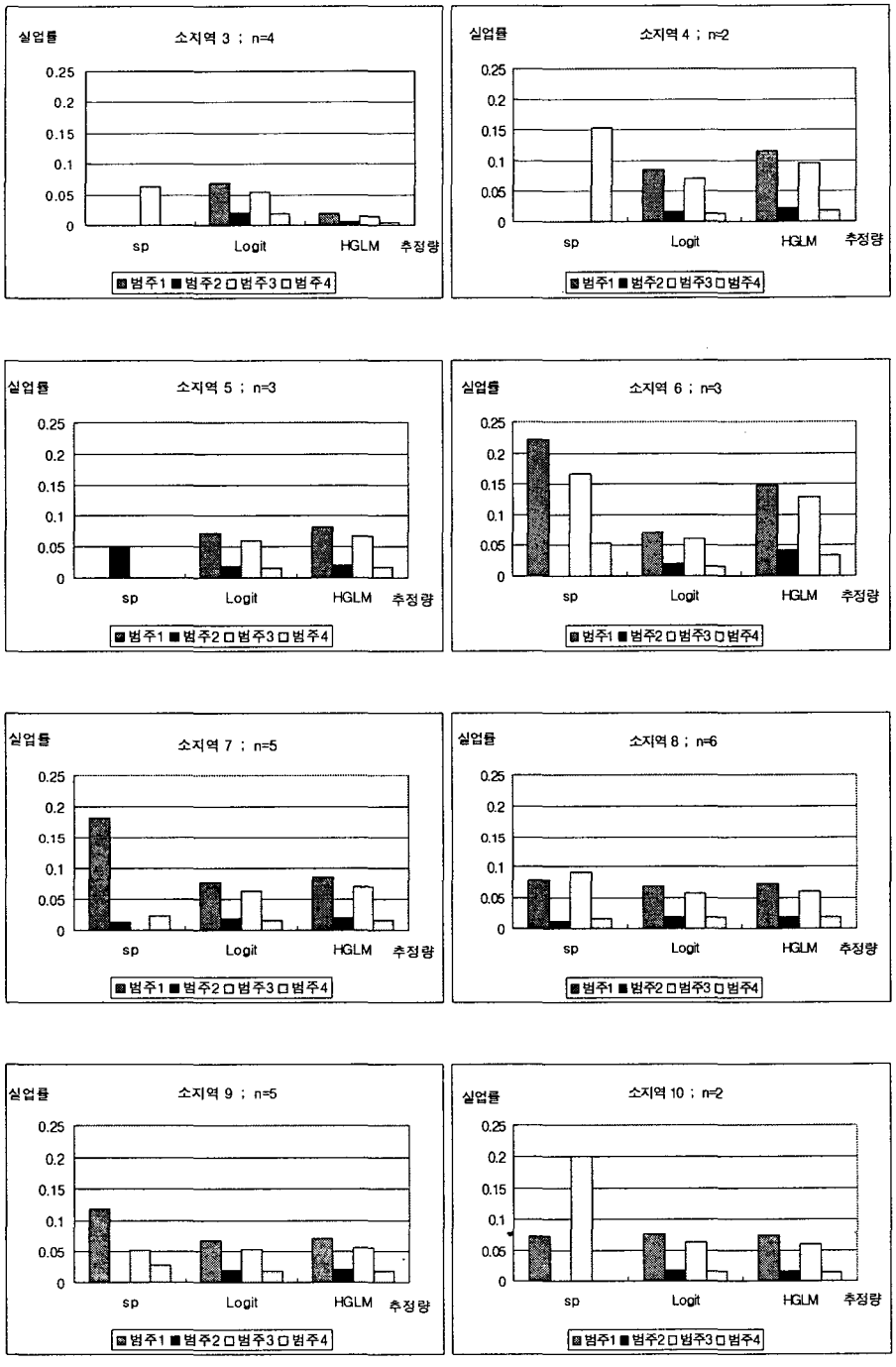
제시된 로지스틱모형과 HGLM을 활용하면 성별, 연령별 요인이 실업률에 주는 영향을 소지역별로 효과적으로 분석할 수 있다. <표 2.1>과 <표 2.2>에 제시된 두 모형에 대한 분석결과를 바탕으로 각 소지역의 범주별(남자=1, 여자=2; 나이(35세 이하)=1, 나이(35세 이상)=2) 실업률 추정결과를 얻을 수 있다. "소지역1"에 대한 성별 및 연령 범주에 따른 실업률을 추정한 결과는 <표 3.1>과 같다. 다른 소지역에 대한 추정결과는 생략하기로 한다.

<표 3.1> "소지역1" 실업률 추정결과 비교

범주	성별	나이	직접	로지스틱	HGLM
1	1	1	0.067227	0.058049	0.059404
2	1	2	0.030303	0.024151	0.024349
3	2	1	0.026316	0.041004	0.041271
4	2	2	0.017751	0.023546	0.023686

이들 추정방법들의 특성 및 차이를 비교하기 위해 충북 10개 시군구 소지역에 대한 성별-연령별 실업률 추정 결과를 그래프로 정리하면 <그림 3.1>과 같다. 여기서 SP, Logit, HGLM은 각각 직접추정량, 로지스틱 합성추정량, HGLM 추정량을 나타낸다.





<그림 3.1> 충청북도 시군구 성별-연령별 소지역 실업률 추정결과 비교

<그림 3.1>을 보면 HGLM 추정량과 로지스틱 추정량은 직접추정량과는 달리 각 소지역에서 전반적으로 범주별 실업률에 있어서 비슷한 패턴을 보여주고 있다. 특히 "소지역3"이나 "소지역4"와 같이 각 소지역의 표본 조사구수가 극히 적은 경우, 이런 현상이 명확하게 나타난다. 실제 이들 소지역의 경우 직접추정 결과 일부 범주에서 실업률은 0인 것으로 산출되지만, 로지스틱 및 HGLM 추정에서는 다른 소지역의 범주별 실업률을 이용한 평활(smoothing) 작업이 수행되기 때문에 이런 결과가 나타난 것이다. 아울러 로지스틱과 HGLM 추정결과를 비교해 보면 로지스틱에 의해 추정된 결과에서는 지역간에 차이가 거의 나타나고 있지 않지만(실제 이 차이는 로지스틱 모형에 변수 Z_{ik}가 포함되어 나타난 것임), HGLM에서는 지역 특성을 나타내는 랜덤효과가 포함되어 있어 지역간의 차이가 추정결과에 반영되어 있다. 이런 관점에서 HGLM은 지역별 효과를 가미한 새로운 복합추정량이라 볼 수 있다. 한편, 본 연구에서는 HGLM에서 지역특성을 나타내는 랜덤효과가 서로 독립적이라고 가정하고 있지만, 인접한 지역의 경우 이들 랜덤효과를 나타내는 변수 u_i가 공간상관관계를 가질 수 있다. 따라서 향후 지역 랜덤효과에 대해 Besag(1974)이 제시한 CAR(conditionnal autoregressive) 등과 같은 공간상관관계를 설명할 수 있는 가정을 추가한 HGLM에 대한 연구가 필요할 것으로 판단된다.

4. 소지역 실업률 추정결과 및 효율성 비교

현행 경제활동인구조사에서 추정대상은 성별과 연령을 구분하지 않은 전체 경제활동인구에 대한 실업률이다. 따라서 제시된 로지스틱 합성추정량 또는 HGLM 추정결과를 토대로 성별, 연령별 추정결과를 통합한 각 소지역에 대한 실업률 산출이 필요하다. 이에 따라 다음과 같이 각 소지역의 성별, 연령별 상주인구를 가중치로 하여 지역별 실업률을 산출한다.

$$\hat{p}_i = \sum_k W_{ik} \hat{p}_{ik} \tag{4.1}$$

여기서, W_{ik}는 i지역에서 k범주의 구성비율로 i지역 실업률 산출을 위한 가중치이고, $\hat{p}_{ik}$은 로지스틱 또는 HGLM에 의한 i지역에서 k범주에 대한 실업률 추정값을 나타낸다.

앞에 제시된 추정방법에 따라 충청북도 10개 시군구 실업률을 추정한 결과와 잭나이프 MSE를 정리한 결과는 <표 4.1>과 같다. 여기에는 식(2.1)에 의한 직접추정량, 식(2.2)의 로지스틱모형을 이용하여 식(4.1)의 과정을 통해 얻은 로지스틱 합성추정량, 그리고 이들 추정결과를 식(2.3)에 반영한 복합추정량에 의한 추정결과가 제시되어 있다. 한편 HGLM의 경우 소지역 추정량의 MSE 추정방법에 대한 이론이 아직 명확하게 규명되어 있지 않고, 잭나이프를 적용하는 방법도 정립되어 있지 않다. 따라서 <표 4.1>에는 HGLM에서 얻은 결과를 식(4.1)에 적용해 얻은 각 소지역별 실업률 추정결과만을 제시하고 있다. HGLM 관련 연구결과는 소지역 추정문제에서 HGLM의 도입 가능성을 점검한 아주 기초적인 수준에 지나지 않으며, 아직 미진한 HGLM 관련 이론연구가 향후 진행됨에 따라 제안된 방법에 대한 보다 심층적인 이론개발이 가능해질 것으로 기대된다. 특히 HGLM에서 소지역 추정량의 MSE 추정 문제는 매우 중요한 연구주제임에도 불구하고, 본 연구에서는 이를 전혀 다루지 못하고 있다는 한계를 갖고 있다.

<표 4.1> 충북지역의 시군구 단위 행정자치구역들의 실업률 및 MSE 추정결과

소지역	표본 조사구 수	직접추정량		합성추정량		복합추정량		HGLM
		$\hat{p}_i^d$	mse_J	$\hat{p}_i^s$	mse_J	$\hat{p}_i^c$	mse_J	$\hat{p}_i^H$
1	22	0.033427	0.000044	0.036357	0.000041	0.034357	0.000023	0.036836
2	11	0.034281	0.000146	0.035407	0.000037	0.034869	0.000015	0.036152
3	4	0.009195	0.000080	0.034887	0.000037	0.019676	0.000008	0.009564
4	2	0.034236	0.001009	0.030259	0.000049	0.030807	0.000002	0.040782
5	3	0.023945	0.000541	0.03363	0.000040	0.031459	0.000004	0.037505
6	3	0.054701	0.000654	0.033669	0.000040	0.037751	0.000007	0.070875
7	5	0.02831	0.000369	0.032791	0.000043	0.031481	0.000006	0.036277
8	6	0.032669	0.000285	0.034349	0.000038	0.033775	0.000012	0.035817
9	5	0.031924	0.000239	0.034973	0.000037	0.033823	0.000008	0.035943
10	2	0.032042	0.000802	0.03271	0.000043	0.032597	0.000004	0.030647

<표 4.1>에 제시된 각 소지역에 대한 직접, 간접, 복합 추정량의 MSE 추정결과를 토대로 좀더 효과적으로 세 추정량의 효율성을 비교하기 위해 다음과 같은 RRMSE(relative root mean square error)를 구하여 정리하면 <표 4.2>와 같다.

$$\mathrm{RRMSE}(\hat{p}_i) = \frac{\sqrt{mse_J(\hat{p}_i)}}{\hat{p}_i} \times 100\,(\%)$$

<표 4.2> 소지역별 추정량들의 RRMSE(%) 비교

소지역	직접추정량	합성추정량	복합추정량
1	20.74431	17.57245	13.85508
2	37.75097	17.13290	11.02243
3	27.98921	17.52660	10.48407
4	99.14218	23.03141	4.62065
5	72.57124	18.91558	5.71668
6	79.80434	18.86353	7.41591
7	59.91334	19.96538	7.66126
8	52.65451	18.06274	10.26760
9	48.20298	17.45166	8.28826
10	88.37587	20.03795	5.73270

<표 4.2>를 보면 복합추정량을 활용함으로써 전반적으로 RRMSE를 대폭 줄일 수 있다는 것을 확인할 수 있다. "소지역1"을 보면 직접추정량과 복합추정량의 RRMSE가 다른 소지역에 비해 상대적으로 크게 차이가 나지 않는다. 이는 "소지역1"의 경우 표본크기가 상대적으로 크기 때문에

직접추정량의 정도(precision)가 크게 떨어지지 않아 효율성에 있어서 복합 추정량에 비해 큰 차이가 나지 않는 것이다. 하지만 표본크기가 작은 다른 소지역의 경우 RRMSE가 크게 차이가 나게 된다. 참고로 본 연구에서 사용된 자료에서 일부 소지역의 경우 표본조사구 중 실업자수가 0인 조사구들이 다수 포함되어 있다. 예를 들어 소지역 3, 4, 5, 6 등의 경우 표본 조사구 중 실업자 수가 0으로 조사된 조사구수가 각각 3, 1, 2, 1개이다. 이런 현상 때문에 직접추정량의 경우 MSE가 합성 또는 복합 추정량의 MSE에 비해 일반적으로 생각하는 것보다 매우 큰 차이를 보이고 있다. 따라서 충분한 크기의 표본 조사구를 확보할 수 없는 소지역에 대한 신뢰할 수 있는 실업률 통계 산출을 위해서는 복합추정량 등 본 연구에서 검토한 소지역 기법들을 적극 활용하는 것이 효과적이라는 것을 다시 한번 확인할 수 있다.

하지만 본 연구에서의 효율성 비교는 제한적인 지역에서 특정 시점의 자료에서 얻은 MSE 추정결과를 토대로 했기 때문에 이런 효율성 비교 결과가 항상 성립된다고 볼 수는 없다. 따라서 보다 명확한 효율성 비교를 위해서는 추가적인 표본 또는 전수조사와 연계한 보다 심층적인 연구가 필요하다. 즉 본 연구에서는 제시된 추정기법 도입에 따른 효율성 향상 가능성을 보여주고 있고, 통계청 등에서 이런 기법을 본격적으로 활용하기 위해서 이에 대한 보다 심도 있는 연구를 수행하는 것이 상당히 의미 있을 수 있다는 것을 시사한다는 것이 정확한 해석일 것이다.

5. 결론 및 제언

본 연구에서는 실업률 추정과 같이 관심변수가 이항분포에 따르는 경우 적용할 수 있는 소지역 추정기법을 제시하고 있으며, 구체적으로 통계청에서 실시되는 경제활동인구조사에서 주 관심 대상인 시군구 실업률을 효율적으로 산출할 수 있는 소지역기법을 개발하는 것을 목적으로 하고 있다. 이를 위해 특히 로지스틱 회귀모형을 소지역추정에 적극 활용하는 방법을 중점적으로 검토하는 동시에 로지스틱 모형에 지역특성을 반영할 수 있도록 랜덤효과를 포함하는 모형을 구현하기 위해 HGLM을 도입하는 방안을 검토하고 있다.

제시된 소지역 추정기법들의 효율성 검증을 위해 2000년 12월 기준 경제활동인구조사 자료를 토대로 충청북도 내의 10개 시군구 단위 행정자치구역의 실업률 추정결과를 산출하고, 이들 추정량의 효율성 비교를 위해 잭나이프 MSE를 계산하여 이를 토대로 구한 RRMSE을 비교하고 있다. 아울러 제시된 로지스틱 모형과 HGLM 모형을 활용하여 성별-연령별 실업률을 산출하는 방안을 제시하고, 추정방법에 따른 특성상의 차이를 검토하고 있다.

제한적이지만 효율성 비교 결과 직접 추정량의 경우 각 시군구에 배정된 조사구가 매우 적어질 수 있기 때문에 신뢰할 수 있는 실업률을 추정하는 데 한계를 갖고 있지만, 제시된 로지스틱 모형을 활용한 복합추정량을 사용하면 직접추정량에 비해 그 효율성을 대폭 향상시킬 수 있다는 것을 알 수 있었다.

한편, 성별-연령별 요인이 실업률에 미치는 영향을 파악하기 위해 로지스틱 회귀모형을 활용한 합성추정법과 지역 랜덤효과가 추가된 HGLM을 활용하는 방법을 검토하였다. 실제 경제활동인구조사에서는 이와 같은 성별-연령별 실업률 분석은 아직 이루어지고 있지 않지만, 향후 제시된 기법을 활용하면 소지역에 따른 실업률 차이를 성별 및 연령에 따른 요인으로 설명하는 것이 가능할 수 있을 것으로 예상된다.

이항 자료를 이용한 소지역 추정을 위해 HGLM을 활용하는 방안은 매우 유용할 것으로 판단되

지만, 아직 필요한 이론이 정립되어 있지 않아 향후 이에 대한 좀더 추가적인 이론 개발이 요구된다. HGLM에서 산출되는 소지역 추정량은 BLUP에서와 마찬가지로 결과적으로 모형에서 얻어진 예측(predictor) 결과를 토대로 산출되게 되는데, 이런 소지역 추정량의 MSE 추정에 대한 이론은 아직 정립되어 있지 않아 이에 대한 연구가 향후 필요하다. 아울러 본 연구에서는 HGLM에서 지역특성을 나타내는 랜덤효과가 서로 독립적이라고 가정하고 있지만, 실제적으로 인접한 지역의 경우 이들 랜덤효과가 공간상관관계를 가지는 것이 보다 일반적이다. 따라서 향후 CAR 등과 같은 공간상관관계를 도입한 HGLM을 기초로 한 소지역 실업률 추정기법에 대한 연구가 필요할 것으로 판단된다.

참고문헌

[1] 김영원, 성나영 (2000). 소지역 통계 생산을 위한 추정 방법. *Journal of the Korean Data and Information Science Society*, 11, 111-126.

[2] 김영원, 조란 (2004). Small Area Estimation of Unemployment rate for the Economically Active Population Survey, *Journal of the Korean Data and Information Science Society*, 15, 1-10.

[3] 박종태, 이상은 (2001). 소지역 추정법에 관한 비교 연구, *Journal of the Korean Data and Information Science Society*, 12, 47-55.

[4] 이계오, 류제복, 김영원 (2001). 소지역 실업률 추정 기법 및 전산 프로그램 개발 보고서, 「노동부 학술연구용역보고서」.

[5] 정연수, 이계오, 이우일 (2003) 시군구 실업자 총계 추정을 위한 설계기반 간접추정법, 「응용통계연구」, 16, 1-14 .

[6] Ambler, R. C., Chambers, D., Kovacevic, M. (2001). Combining Unemployment Benefits Data and LFS Data to Estimate ILO Unemployment for Small Areas; An Application of a Modified Fay-Herriot Method, *Bulletin of the ISI 2001*, 128-138.

[7] Besag, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems, *Journal of the Royal Statistical Society, Series B,* 36, 192-236.

[8] Chung, Y. S., Lee K-O., and Kim, B. C. (2003). Adjustment of Unemployment Estimates Based on Small Area Estimation in Korea, *Survey Methodology,* 29, 45-52.

[9] Drew, 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.

[10] Ghosh, M. and Rao, J. N. K. (1994). Small Area Estimation: An Appraisal, *Statistical Science,* 9, 55-93.

[11] Ghosh, M., Natarajan, K., Stroud, T. W. F. and Carlin, B. P. (1998). Generalized Linear Models for Small-Area Estimation. *Journal of American Statistical Association,* 93, 273-282.

[12] Jiang, J., Lahiri, P., Wan, S_M., Wu, C-H. (2001). Jackknifing in the Fay-Herriot Model with an Example. Technical Report, Department of Statistics, University of Nebraska.

[13] Lee, Y. and Nelder, J. A. (1996). Hierarchical Generalized Linear Models (with discussion).

Journal of the Royal Statistical Society, B 58, 619-78.

[14] Lee, Y. and Nelder, J. A. (2001). Hierarchical Generalized Linear Models: A Synthesis of Generalized Linear Models, Random-Effect Models and Structured Dispersions. *Biometrika,* 88, 987-1006.

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

[2004년 5월 접수, 2004년 10월 채택]