응용통계연구 (2010)
**23**(5), 891–908
DOI: 10.5351/KJAS.2010.23.5.

# 소지역 실업자수 추정을 위한 로지스틱 선형혼합모형 기반 EBLUP 타입 추정량 평가

김서영[1] · 권순필[2]

[1]통계개발원 조사연구실, [2]통계개발원 조사연구실

(2010년 4월 접수, 2010년 8월 채택)

**요 약**

근래 소지역 추정(small area estimation)에 관한 연구는 비교적 활발하게 이루어진 편인데 비해, 우리나라의 국가통계 작성에 실제 활용된 사례는 거의 없는 실정이다. 이는 소지역 추정이 갖는 많은 장점에도 불구하고 공식통계 활용 여부를 판단하기가 그만큼 어렵기 때문이다. 본 연구는 소지역 추정방법에 의해 우리나라 시군구 실업자 통계를 생산하는 방법을 모색하고자 한다. 시군구 실업자수 추정은 로지스틱 선형혼합모형에 의한 EBLUP 타입(EBLUP-type) 추정량을 사용하였다. 실제자료분석과 모의실험 결과에 대해 다양한 평가 방법을 적용하고, 추정량의 특성을 비교 분석하였다. 그 결과 본 연구에서 적용한 로지스틱 선형혼합모형 기반 EBLUP 타입 추정량은 우리나라 시군구 실업자수 추정에 활용 가능성이 높은 것으로 평가되었다.

주요용어: 소지역 추정, 실업자, 일반선형혼합모형.

## 1. 서론

최근 사용되고 있는 다양한 통계작성 방법 중 조사에 의한 방법이 차지하는 비중이 높은 편이다. 하지만 사회가 복잡해지고 조사환경이 어려워질수록, 조사비용과 오차는 통계생산 기관의 큰 부담이며, 이러한 부담은 소지역 통계일수록 클 수밖에 없다. 이에 대해 많은 국가들은 행정자료를 잘 정비하여 통계목적으로 활용한다거나 모형을 이용한 추정방법 등을 활용함으로써, 조사통계에 대한 부담을 극복하고자 하였다. 특히, 소지역 통계 작성에 있어서는 미국, 영국, 호주 등 통계선진국가를 중심으로 소지역 추정 방법이 다양한 분야에서 널리 사용되고 있다. 실업은 일반적인 사회경제 상황을 측정할 수 있는 지표로서, 이미 오래전부터 우리 사회의 주요 관심사가 되었다. 이와 함께 실업을 비롯한 고용통계는 소지역, 지방 및 중앙 정부가 고용정책을 수행하는데 필요한 자금을 효과적으로 배정하는데 필요한 핵심 정보가 되었다. 유럽 공동체는 다른 유럽 국가들과의 상호 균등 발전을 위해 특별 고용 프로그램을 지원하는 자금을 제공하고 있고, 이러한 자금 배정은 소지역의 신뢰할만한 통계에 근거하여 이루어지고 있다 (Molina 등, 2007). 실제로 많은 국가들에서 그러했듯이, 광범위한 유럽 국가 내 작은 지역들의 고용통계를 조사에 의해 파악한다는 것은 현실적으로 어려운 일일 것이다. 왜냐하면 소지역에서 신뢰할 만한 조사통계를 작성하려면 우선 표본크기가 커야 하고, 그에 따른 많은 인력과 예산뿐만 아니라 이로부터 파생되는 조사기간의 연장, 비표본오차의 증가 등 현실적인 문제가 발생할 수 있기 때문이다.

우리나라도 지역 특성을 고려한 보다 세분화된 고용 정보에 대한 수요가 2000년을 전후로 증가하였다. 경기도를 비롯한 일부 지역은 각 지자체에서 해당 시군 지역의 고용현황을 파악하기 위해 조사를 실시

[1]교신저자: (302-724) 대전시 서구 월평동 282-1 나라키움 통계센터, 통계개발원 조사연구실, 사무관.
E-mail: smilegong@korea.kr

하고 있다. 통계청은 2008년에 정부의 필요에 의해 우리나라 모든 시군에 대해 '지역별 고용조사'를 실시하였다. 이처럼 소지역의 고용정보 수집은 많은 예산과 조사 부담에도 불구하고 그 수요는 날로 커지고 있는 현실이다. 경제활동인구조사(경활조사)는 전국 또는 시도 단위 통계 작성을 목적으로 설계되었기 때문에 소지역에서의 표본이 작고, 상대표준오차(relative standard error) 또는 CV(Coefficient of variation)가 커서 소지역 통계로서의 신뢰도는 매우 낮아질 수 있다. 이처럼 표본설계시에 고려되지 않은 시군구에서 의미 있는 직접추정치(direct estimate)는 사실상 구하기가 쉽지 않다.

우리나라에서 소지역 추정 연구는 2000년을 전후로 관심을 받기 시작하였다. 이계오 (2000), 박종태와 이상은 (2001), 정연수 등 (2003)은 설계기반 소지역 추정방법들을 중심으로 그 특성을 비교하고 분석하였다. 그 이후 국제적 흐름에 따라 모형 추정량, 베이지안 추정량 및 보조정보 활용과 같은 방법론 연구 (이상은, 2006; 김정숙 등, 2008, 김서영과 권순필, 2010)에 대한 관심이 높아졌으며, 실무적 활용 차원에서의 연구도 증가하게 되었다 (김수택 등, 2008; 김서영과 권순필, 2009).

다른 나라들의 소지역 추정 활용사례를 국가통계의 입장에서 보면, 그 방법은 그렇게 복잡하거나 다양하지는 않은 것 같다. 고용통계에 있어서 미국과 일본은 시계열 모형 (Tiller, 1992), 캐나다는 횡종단 모형 (Rao와 Yu, 1994), 영국은 로지스틱 선형혼합모형 (ONS, 2009)을 사용하고, 호주는 영국과 동일한 모형을 검토하고 있다. 우리나라도 위의 방법들을 중심으로 검토하고 일반선형모형을 경제활동인구조사에 적용하여 연구를 진행하고 있다 (김서영과 권순필, 2009).

본 연구는 소지역 추정방법의 활용성 측면, 특히 국가통계 작성에서의 소지역 추정을 접근하고자 하였다. 특히 고용통계 중 시군구 실업자 총수 추정을 목적으로 하였다. 이를 위해 로지스틱 선형혼합모형을 이용한 EBLUP(Empiricla Best Linear Unbiased Predictor; EBLUP) 타입 추정 방법을 사용하였다 (Saei와 Chambers, 2003a). 모형에서 모수는 MPQL(Maximum penalized quasi likelihood), 분산성분은 REML(Restricted maximum likelihood) 방법으로 추정하였다. 추정결과에 대해 모형의 적합성, 추정치의 정확성 및 신뢰성 측면에서 5개의 평가측도를 사용하였다. 또한 2005년 센서스 자료를 이용한 모의실험을 통해 추정치의 편향을 검증하였고, 대표본조사 결과와의 비교를 통해 모형 추정치의 타당성 및 신뢰성을 확인하였다. 또한 다른 추정량들과의 상대적 우위성 평가도 동시에 수행하였다.

본 논문의 구성은 다음과 같다. 2절에서는 로지스틱 선형혼합모형과 EBLUP 타입 추정량에 대해 설명하였다. 3절에서는 분석에 사용된 자료, 모형, 추정, 평가 방법들을 설명하였다. 4절에서는 실제자료분석과 모의실험 결과를 통해 소지역 추정 방법의 활용 가능성을 평가하고, 마지막으로 결론과 향후과제를 언급하였다.

## 2. 로지스틱 선형혼합모형에 기반한 EBLUP 타입 추정량

### 2.1. 이항반응자료에 적용

$\mathbf{y}_s = \{y_{sdi}\}$, $\mathbf{y}_r = \{y_{rdi}\}$가 각각 이항반응변수의 표본과 비표본값에 해당하는 벡터라고 하자. 즉, 실업자 추정의 경우, $y_{sdi}$는 $d$지역 $i$그룹 내에서 표본으로부터 관측된 실업자 수이고, $y_{rdi}$는 모집단 가운데 표본으로 추출된 나머지(비표본)에 포함된 실업자수라 하자. 관심 모수가 $\theta = \mathbf{a}_s\boldsymbol{y}_s + \mathbf{a}_r\boldsymbol{y}_r$일 때, $p_{di}$는 소지역 $d$의 성×연령별 그룹 $i$내에서 실업자가 발생할 확률, $N_{di}(n_{di})$는 모집단(표본)에서 소지역 $d$의 그룹 $i$에 포함된 총 개체수라 하자. 이때 $y_{sdi}(y_{rdi})$는 지역내의 성×연령 그룹에 실업이 발생할 확률 $p_{di}$와 평균 $n_{di}p_{di}((N_{di}-n_{di})p_{di})$인 이항분포를 따르는 반응변수이다. 실업이 발생할 확률 $p_{di}$는 식 (2.1)의 로지스틱 선형혼합모형을 따른다 (Saei와 Chambers, 2003a).

$$\eta_{di} = \text{logit}(p_{di}) = \ln\left(\frac{p_{di}}{1-p_{di}}\right) = \mathbf{x}_{di}'\boldsymbol{\beta} + u_d, \tag{2.1}$$

여기서 $\mathbf{x}_{di}$는 지역 내의 성 × 연령별 보조변수 벡터이고, $u_d$는 지역랜덤효과, $\boldsymbol{\beta}$는 회귀계수 벡터이다.

식 (2.1)에서 $\eta$와 $\boldsymbol{u}$에 대한 $\boldsymbol{y}$의 조건부 평균, $E(\boldsymbol{y}|\boldsymbol{u}) = h(\boldsymbol{\eta})$는 각각 $\eta = [\boldsymbol{\eta}'_s, \boldsymbol{\eta}'_r]'$, $E(\boldsymbol{y}|\boldsymbol{u}) = [h'(\boldsymbol{\eta}_s), h'(\boldsymbol{\eta}_r)]'$이고, 여기서 $\eta_s = \mathbf{X}_s\boldsymbol{\beta} + \mathbf{Z}_s\mathbf{u}$, $\eta_r = \mathbf{X}_r\boldsymbol{\beta} + \mathbf{Z}_r\mathbf{u}$, $E(\boldsymbol{y}_s|\mathbf{u}) = h(\boldsymbol{\eta}_s)$, $E(\boldsymbol{y}_r|\mathbf{u}) = h(\boldsymbol{\eta}_r)$과 같다. 이때 행렬 $\mathbf{Z}_s = \mathbf{Z} = \mathrm{diag}(\mathbf{1}_I;\ d = 1, \ldots, D)$, $\mathbf{1}_I$는 길이가 $I$, 모든 원소가 1인 벡터를 나타내고, $\mathbf{X}_s = \mathbf{X} = [\mathbf{x}_{11} \cdots \mathbf{x}_{1I} \cdots \mathbf{x}_{d1} \cdots \mathbf{x}_{di} \cdots \mathbf{x}_{DI} \cdots \mathbf{x}_{DI}]'$이다. 이항반응변수 벡터 $\mathbf{y}_s$, $\mathbf{y}_r$의 $\mathbf{u}$에 대한 조건부 평균은 각각 식 (2.2), (2.3)과 같이 표현될 수 있다. 이때, $h$는 링크함수(link function)의 역함수로서 이항분포를 가정한 로짓 링크함수를 나타낸다.

$$E(\mathbf{y}_s|\mathbf{u}) = h(\boldsymbol{\eta}_s) = n_{di}\left(\frac{\exp(\mathbf{X}_s\beta + \mathbf{Z}_s\mathbf{u})}{1 + \exp(\mathbf{X}_s\beta + \mathbf{Z}_s\mathbf{u})}\right), \tag{2.2}$$

$$E(\mathbf{y}_r|\mathbf{u}) = h(\boldsymbol{\eta}_r) = (N_{di} - n_{di})\left(\frac{\exp(\mathbf{X}_r\beta + \mathbf{Z}_r\mathbf{u})}{1 + \exp(\mathbf{X}_r\beta + \mathbf{Z}_r\mathbf{u})}\right). \tag{2.3}$$

$\mathbf{u}$에 대한 $\mathbf{y}_s$의 로그우도함수와 $\mathbf{u}$의 확률밀도함수의 로그값은 각각 다음과 같다.

$$l_1 = \text{constant} + \sum_{d=1}^{D}\sum_{i=1}^{I}\left[y_{di}\eta_{sdi} - n_{di}\ln(1 + \exp(\eta_{sdi})\right],$$

$$l_2 = -\frac{1}{2}\left[\text{constant} + D\ln\varphi + \varphi^{-1}\mathbf{u}'\mathbf{u}\right],$$

여기서 $\mathbf{u} = [\mathrm{u}_1, \mathrm{u}_2, \ldots, \mathrm{u}_D]'$는 평균 0이고, 분산 $\varphi\mathrm{I}_D$인 정규분포를 따른다. $\varphi$가 알려졌다는 전제하에, $\boldsymbol{\beta}, \mathbf{u}$에 대해 $l = l_1 + l_2$를 최대로 하는 $\boldsymbol{\beta}, \mathbf{u}$의 추정치를 PL(penalized likelihood)추정치라 부른다 (Saei와 McGilchrist, 1998). 모수 $\theta$를 $\boldsymbol{\beta}, \mathbf{u}$의 PL 추정치로 대체하고, 이것을 $\boldsymbol{\theta}$에 대한 MPQL(Maximum Penealized Quasi Likelihood) 또는 BLUP 타입 PL 추정치라 한다 (식 (2.4)). 이와 같은 추정 절차를 따르면, $\boldsymbol{\theta}$의 최종 추정량은 식 (2.4)와 같다. $\mathbf{u}$의 분산인 $\varphi$가 알려졌다는 전제하에 $\boldsymbol{\beta}, \mathbf{u}$의 자세한 PL 추정 절차는 Saei와 Chambers (2003a)를 참고할 수 있다.

$$\hat{\theta} = \mathbf{a}_s\boldsymbol{y}_s + \mathbf{a}_r\hat{\boldsymbol{y}}_r = \mathbf{a}_s\boldsymbol{y}_s + \mathbf{a}_r h(\hat{\eta}_r) = \mathbf{a}_s\boldsymbol{y}_s + \mathbf{a}_r h(\boldsymbol{X}_r\hat{\boldsymbol{\beta}} + \boldsymbol{Z}_r\hat{\mathbf{u}}). \tag{2.4}$$

### 2.2. REML 추정과 EBLUP 타입 추정량

$\varphi$는 모르는 경우가 많기 때문에 자료로부터 추정해야 한다. $\varphi$에 대한 REML 추정치를 구해 보자. 행렬 $\mathbf{V}$의 분할행렬과 $\mathbf{V}$의 역행렬은 다음과 같다.

$$\mathbf{V} = \begin{bmatrix} \mathbf{X}'_s\mathbf{B}_s\mathbf{X} & \mathbf{X}'_s\mathbf{B}_s\mathbf{Z} \\ \mathbf{Z}'_s\mathbf{B}_s\mathbf{X} & \varphi^{-1}\mathrm{I}_D + \mathbf{Z}'_s\mathbf{B}_s\mathbf{Z} \end{bmatrix}, \qquad \mathbf{V}^{-1} = \begin{bmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{T}_{21} & \mathbf{T}_{22} \end{bmatrix}.$$

행렬의 각 원소에 해당하는 행렬 $\mathbf{T}$는 다음과 같다.

$$\mathbf{T}^*_s = (\varphi^{-1}\mathbf{I}_D + \mathbf{Z}'_s\mathbf{B}_s\mathbf{Z}_s)^{-1}, \qquad \mathbf{T}_{11} = (\mathbf{X}'_s\mathbf{B}_s\mathbf{X}_s - \mathbf{X}'_s\mathbf{B}_s\mathbf{Z}_s\mathbf{T}^*_s\mathbf{Z}'_s\mathbf{B}_s\mathbf{X}_s)^{-1},$$
$$\mathbf{T}_{22} = \mathbf{T}^*_s + \mathbf{T}^*_s\mathbf{Z}'_s\mathbf{B}_s\mathbf{X}_s\mathbf{T}_{11}\mathbf{X}'_s\mathbf{B}'_s\mathbf{Z}_s\mathbf{T}^*_s, \qquad \mathbf{T}_{12} = -\mathbf{T}_{11}\mathbf{X}'_s\mathbf{B}'_s\mathbf{Z}_s\mathbf{T}^*,$$
$$\varphi = D^{-1}(\mathrm{tr}(\mathbf{T}^*_s) + \mathbf{u}'\mathbf{u}),$$

$\boldsymbol{\beta}$, $\mathbf{u}$의 PL 추정과 $\varphi$의 REML 추정에 의해, $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{u}}$, $\hat{\varphi}$이 수렴하면 $\theta$의 EBLUP 타입 추정치는 다음과 같이 표현된다 (Saei와 Chambers, 2003b).

$$\hat{\theta} = \mathbf{a}_s\mathbf{y}_s + \mathbf{a}_r\hat{\mathbf{y}}_r, \qquad \hat{\mathbf{y}}_r = (N_{di} - n_{di})\frac{\exp(x'_{di}\hat{\boldsymbol{\beta}} + \hat{u}_d)}{1 + \exp(x'_{di}\hat{\boldsymbol{\beta}} + \hat{u}_d)}. \tag{2.5}$$

$\hat{\theta}$의 평균 교차곱 행렬(mean cross product matrix)의 추정치는 식 (2.6)과 같다.

$$\widehat{\mathrm{mcpe}}(\hat{\theta}) = G_1(\hat{\varphi}) + G_2(\hat{\varphi}) + 2G_3(\hat{\varphi}) + G_4(\cdot) \tag{2.6}$$

이때,

$$G_1(\hat{\varphi}) = \mathbf{Z}_r^+ \hat{\mathbf{T}}_s^* \mathbf{Z}_R^{+\prime}, \qquad G_2(\hat{\varphi}) = \left[\mathbf{X}_r^+ - \mathbf{Z}_r^+ \hat{\mathbf{T}}_s^* \mathbf{Z}_s' \mathbf{B}_s \mathbf{X}_s\right] \mathbf{T}_{11} \left[\mathbf{X}_r^{+\prime} - \mathbf{X}'_s \hat{\mathbf{B}}_s \mathbf{Z}_s \hat{\mathbf{T}}_s^* \mathbf{Z}_r^{+\prime}\right],$$

$$G_3(\widehat{\varphi}) = \left[\mathrm{tr}\left(\widehat{\nabla}_i \widehat{\mathbf{\Sigma}}_s^+ \widehat{\nabla}_j'\right) \widehat{\mathrm{var}}(\hat{\varphi})\right], \qquad G_4(\cdot) = \mathbf{a}_r \mathbf{B}_r \mathbf{a}_r',$$

$$\mathbf{H}_r = \mathrm{diag}((N_{di} - n_{di})\hat{p}_{di}(1-\hat{p}_{di})), \qquad \mathbf{X}_r^+ = \mathbf{a}_r \mathbf{H}_r \mathbf{X},$$

$$r_1 = \hat{\varphi}^{-2}\mathrm{tr}\left(\hat{\mathbf{T}}_{22}\right), \qquad r_{11} = \mathrm{tr}\left(\hat{\mathbf{T}}_{22}\hat{\mathbf{T}}_{22}\right),$$

$$\widehat{\mathrm{var}}(\hat{\varphi}) = 2\left(\hat{\varphi}^{-2}(D - 2r_1) + \hat{\varphi}^{-4} r_{11}\right)^{-1}, \qquad \widehat{\mathbf{\nabla}}_i = \hat{\varphi}^{-2} \mathbf{Z}_r^+ \hat{\mathbf{T}}_s^* \hat{\mathbf{T}}_s^*, \qquad \mathbf{Z}_r^+ = \mathbf{a}_r \mathbf{H}_r \mathbf{Z},$$

$$\widehat{\Sigma}_s^+ = \mathbf{Z}_s' \hat{\mathbf{B}}_s \mathbf{Z}_s + \hat{\varphi} \mathbf{Z}_s' \hat{\mathbf{B}}_s \mathbf{Z}_s \mathbf{Z}_s' \hat{\mathbf{B}}_s \mathbf{Z}_s,$$

$$\mathbf{B}_s = -\partial^2 l_1 / \partial \hat{\boldsymbol{\eta}}_s \partial \hat{\boldsymbol{\eta}}_s{}' = \mathrm{diag}(n_{di}\hat{p}_{sdi}(1-\hat{p}_{sdi})),$$

$$\mathbf{B}_r = -\partial^2 l_1 / \partial \hat{\boldsymbol{\eta}}_r \partial \hat{\boldsymbol{\eta}}_r{}' = \mathrm{diag}((N_{di} - n_{di})\hat{p}_{rdi}(1-\hat{p}_{rdi})).$$

EBLUP 추정량을 이용하여 우리나라 시군구별 실업자수 총수를 추정할 경우, $d$지역의 실업자 총수 추정량 $\hat{\theta}_d$은 식 (2.7)과 같다.

$$\hat{\theta}_d = \sum_{i=1}^{I} \hat{\theta}_{di} = \sum_{i=1}^{I} \{y_{di} + (N_{di} - n_{di})\hat{p}_{di}\} \tag{2.7}$$

$$\hat{p}_{di} = \mathrm{antilog}(\mathbf{x}_{di}'\hat{\boldsymbol{\beta}} + u_d) = \left[\frac{\exp(\mathbf{x}_{di}'\hat{\boldsymbol{\beta}} + \hat{u}_d)}{1 + \exp(\mathbf{x}_{di}'\hat{\boldsymbol{\beta}} + \hat{u}_d)}\right]$$

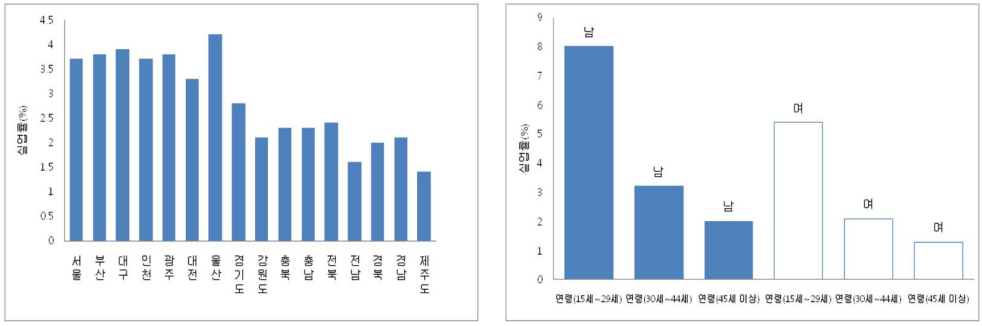
이때, $y_{di}$은 소지역 $d$에서 성×연령 그룹 $i$에 속한 표본에서 조사된 실업자 수, $n_{di}$은 소지역 $d$의 성×연령 그룹 $i$에서 추출된 표본 수, $N_{di}$는 소지역 $d$의 성×연령 그룹 $i$에 포함된 15세 이상 경제활동인구의 추계인구를 나타낸다.

## 3. 경제활동인구조사를 이용한 시군구 실업자수 추정

로지스틱 선형혼합모형 기반 EBLUP 타입 추정량을 이용하여 우리나라 시군구 실업자 총수를 추정해 보자. 종속변수는 경활조사의 실업자수, 보조변수는 고용보험자료의 실업급여등록자수와 인구·지리적 특성을 반영한 더미변수를 사용한다. 소지역은 230개 모든 시군구를 대상으로 하고, 자료는 2007년 11월부터 2008년 10월까지의 자료를 사용했다. 대표본조사 추정치와의 비교 시점은 2008년 10월이다.

### 3.1. 자료

경활조사는 약 32,000 가구의 15세 이상 인구 약 72,000여명을 대상으로 (통계청, 2008), 매월 우리나라의 고용현황을 파악하는 조사이다. 이 조사는 전국과 16개 시도 통계 작성을 목적으로 설계되었으며, 시군구 지역은 표본설계시에 고려되지 않는다. 따라서 소지역에 할당된 표본크기는 매우 작게 된다. 경활조사의 표본추출은 조사구(약 60 가구 포함)를 추출한 후, 조사구내 가구를 추출하는 방식으로 이루어진다. 추출방법은 우선 2005년 인구주택총조사에 의해 설정된 조사구 중에서 1629개 조사구를 추출한다. 표본가구는 각 조사구를 평균 5가구씩 조사구역으로 분할한 후, 표본으로 뽑힌 1개 구역과 북쪽,




(a) 시도별 (b) 성×연령별

**그림 3.1.** 16개 시도별과 성×연령별 실업률

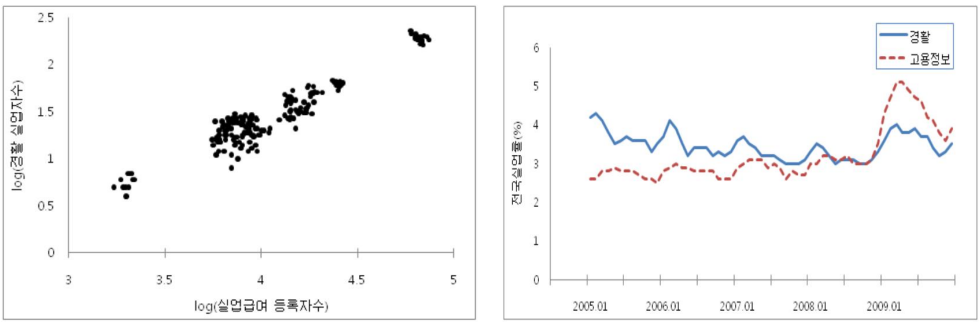
시계방향으로 인접한 3개 구역을 추출한다. 즉, 4개 구역의 평균 20 가구를 추출한다 (KOSIS, 2010). 2008년 당시 우리나라 소지역 개수는 232개로, 표본조사구수가 0개인 지역은 옹진군, 울릉군, 신안군 등의 섬지역이 해당한다. 2008년 10월 경활조사자료의 소지역별 표본조사구에 대한 기초 통계량은 아래와 같다.

| | 최소값 | 1사분위수 | 중위수 | 평균 | 3사분위수 | 최대값 |
|---|---|---|---|---|---|---|
| 조사구수 | 0 | 2 | 4 | 5.8 | 6 | 32 |

실업급여등록자수는 우리나라 임금근로자들을 대상으로 실업급여수급자 현황을 파악할 수 있는 한국고용정보원의 행정등록자료이다. 실업자수 추정을 위한 모형은 각 시군구내의 6개 성별(남자/여자), 연령별(30세 미만/30세 이상 45세 미만/45세 이상)범주의 실업인구를 개체로 사용한다. 모형에는 성×연령별 그룹의 실업급여등록율, 소지역의 실업급여등록율, 성×연령별 그룹(6개), 행정구역상 시도 그룹(16개 시도), 시군 그룹(2개)과 소지역의 실업급여등록률에 대해 성, 연령별 교호작용을 고려하였다.

그림 3.1은 2008년 10월 경활조사의 16개 시도와 성×연령별 실업률을 나타낸 것이다. (a)는 우리나라 특광역시와 시도 간에는 실업률 차이를 뚜렷하게 보여준다. 광역시의 경우 대전과 울산의 실업률 차이가 가장 크고, 시도 실업률의 경우는 전남이 가장 낮고 경기가 가장 높은 등 지역 간 차이가 크게 나타난다. (b)는 성이 같을 경우 연령별 실업률에 큰 차이가 있고, 20세 미만 연령대에서는 남녀 간의 실업률 차이가 있음을 보여준다. 본 연구는 이러한 각 그룹 내의 실업률의 차이를 모형에 반영함으로써 추정치의 정도를 높이고자 하였다.

그림 3.2는 모형에서 보조변수로 사용한 실업급여등록자수와 경활조사 실업자간의 관계를 나타낸다. 그림 3.2의 (a)는 2008년 10월 시점의 전국 수준에서 실업급여등록자수와 경활조사 실업자수의 산점도로서, 두 변수 간의 피어슨 상관계수는 0.96으로 매우 높은 상관성을 갖는다. 그림 3.2의 (b)는 최근 3년간(2006년~2009년)간의 고용정보원 전국 실업률(= 실업급여등록률: 실업급여등록자수/전체 피보험자수)과 경활조사 전국 실업률(= 실업자수/15세 이상 경제활동 인구수)의 시계열을 나타낸 것으로, 두 실업률간의 전체적 경향이 비슷하다는 것을 알 수 있다. 2009년에 접어들면서 고용정보원 실업률이 급격하게 증가한 것은 해당 시점에서 당시 실직한 임금근로자들의 실업급여청구가 늘어났기 때문으로 해석될 수 있다. 또한 두 실업률 간에 발생하는 약간의 수치적 차이는 조사대상에 대한 커버리지 차이라 할 수 있다. 즉, 경활조사의 실업자수는 우리나라 15세 이상 인구를 대상으로 하는 반면, 실업급여등록자수는 임금근로자만을 대상으로 한다는 점에서 조사 특성상 차이가 있기 때문이다.




(a) 실업급여등록자수와 경활실업자수의 산점도 (b) 실업급여등록률과 경활 실업률의 시계열도

**그림 3.2.** 실업급여등록자수와 경활실업자수의 산점도(a), 최근 3년(2006년~2009년) 실업급여등록률과 경활 실업률의 시계열도(b)

### 3.2. 모형적합

$p_{di}$는 표본에서 $d$지역의 성×연령 범주 $i$에서 조사된 개체들이 실업일 확률이라 하자. 이항반응자료에 대한 소지역 추정 모형을 위해 식 (3.1)의 로지스틱 선형혼합모형을 가정한다.

$$\text{logit}(p_{di}) = \ln\left(\frac{p_{di}}{1-p_{di}}\right) = \mathbf{X}_{di}\boldsymbol{\beta} + u_d, \quad d = 1,\ldots,230,\ i = 1,\ldots,6,\ u_d \sim N(0,\varphi). \tag{3.1}$$

식 (3.1)에서 $\mathbf{X}_{di}$는 소지역 $d$에서 성×연령에 대한 보조변수 벡터이고, $u_d$는 평균 0, 분산 $\varphi$인 정규분포를 따르는 지역 랜덤효과이다. 식 (3.1)을 보조변수에 대해 구체적으로 표현하면 식 (3.2)와 같다.

$$\begin{aligned}\text{logit}(p_{di}) = {} & \beta_0 + \beta_1\text{sex} + \beta_2\text{age1} + \beta_3\text{age2} + \beta_4\text{sexage1} + \beta_5\text{sexage2} \\ & + \beta_6\text{sido1} + \cdots + \beta_{20}\text{sido15} \\ & + \beta_{21}\text{logit}(\text{실업급여등록율}_{di}) + \beta_{22}\text{logit}(\text{실업급여등록율}_{di})\text{sex} \\ & + \beta_{23}\text{logit}(\text{실업급여등록율}_{di})\text{age1} + \beta_{24}\text{logit}(\text{실업급여등록율}_{di})\text{age2} \\ & + \beta_{25}\text{logit}(\text{실업급여등록율}_{di})\text{sexage1} + \beta_{26}\text{logit}(\text{실업급여등록율}_{di})\text{sexage2} \\ & + \beta_{27}\text{logit}(\text{실업급여등록율}_{d}) + \beta_{28}\text{sigun} + u_d.\end{aligned} \tag{3.2}$$

모형에서 각 모수 추정은 $\boldsymbol{\beta}$와 $\boldsymbol{u}$의 MPQL 추정과 $\varphi$의 REML 추정을 결합한 반복절차를 사용하였다 (Saei와 Chambers, 2003a). MPQL 추정과 REML 추정은 각 모수에 대한 초기값과 수렴값으로 $\varphi_0 = 0.1$, $\beta_0 =$ GLM 추정치, $u_0 = 0$, 수렴값 $= 0.0001$을 사용하였고, 100번 반복 수행하였다. 실제 계산은 R 프로그램을 사용하였다.

표 3.1은 실제 자료를 식 (3.2)의 모형에 적합한 결과이다. 표 3.1의 주석은 $\varphi$에 대한 추정치와 회귀계수 검정을 위한 유의수준과 관련된 정보를 나타낸다. 표 3.1에서 sex(남자)는 유의수준 0.1에서 유의하고 남자의 실업 증가율은 여자에 비해 약 15배 (exp(2.685))높다. sido변수의 경우 기준범주는 제주도 (sido = 16)이고 sido = 1부터 sido = 7은 유의수준 0.001, 0.01, 0.05 수준에서 각각 유의한 것으로 나타났다. 이 지역들은 모두 광역시 단위로서 제주도에 비해 실업 증가율이 최소 약 1.7배에서 최대 2.5배 정도 높게 나타났다. sido = 8부터 sido = 15까지는 모두 도 단위 지역으로서 제주도에 대해 실업 증가율이 통계적으로 유의하지는 않았다. 나머지 변수들에 대해서도 유사한 방법으로 설명될 수 있다. 모형적합과 관련하여 AIC(Akaike Information Criterior) 또는 $R^2$ 관련 통계량 기준에서 가장 적합한 모형을 최종 모형을 선택하였다. 이에 대한 결과는 지면 한계상 자세하게 설명하지 않기로 한다.

표 **3.1.** 실업에 대한 모형적합 결과

| 변수 | 추정치 | 표준편차 | $Z$ 통계량 | $p$값 |
|---|---|---|---|---|
| 상수 | −4.870 | 1.364 | −3.57 | 0*** |
| sex | 2.685 | 1.420 | 1.89 | 0.058§ |
| age = 1 | 2.006 | 1.513 | 1.32 | 0.184 |
| age = 2 | −0.337 | 1.643 | −0.20 | 0.837 |
| sexage = 1 | −0.727 | 1.994 | −0.36 | 0.715 |
| sexage = 2 | −1.486 | 2.210 | −0.67 | 0.501 |
| sido = 1 | 0.677 | 0.262 | 2.58 | 0.009** |
| sido = 2 | 0.588 | 0.285 | 2.05 | 0.039* |
| sido = 3 | 0.619 | 0.281 | 2.20 | 0.027* |
| sido = 4 | 0.616 | 0.279 | 2.20 | 0.027* |
| sido = 5 | 0.685 | 0.277 | 2.47 | 0.013* |
| sido = 6 | 0.521 | 0.281 | 1.85 | 0.063§ |
| sido = 7 | 0.914 | 0.273 | 3.34 | 0*** |
| sido = 8 | 0.338 | 0.266 | 1.27 | 0.203 |
| sido = 9 | 0.201 | 0.306 | 0.65 | 0.510 |
| sido = 10 | 0.399 | 0.296 | 1.34 | 0.177 |
| sido = 11 | 0.300 | 0.306 | 0.99 | 0.321 |
| sido = 12 | 0.303 | 0.291 | 1.04 | 0.298 |
| sido = 13 | 0.559 | 0.285 | 1.95 | 0.050* |
| sido = 14 | 0.365 | 0.297 | 1.22 | 0.219 |
| sido = 15 | 0.267 | 0.291 | 0.91 | 0.359 |
| sex ∗ xdi | −0.612 | 0.168 | −3.64 | 0*** |
| age ∗ xdi = 1 | 0.616 | 0.371 | 1.65 | 0.097§ |
| age ∗ xdi = 2 | 0.757 | 0.451 | 1.67 | 0.093§ |
| sexage ∗ xdi = 1 | 0.390 | 0.547 | 0.71 | 0.475 |
| sexage ∗ xdi = 2 | −0.353 | 0.554 | −0.63 | 0.524 |
| xdi | −0.612 | 0.698 | −0.87 | 0.380 |
| xd | −0.353 | 0.298 | −1.18 | 0.235 |
| sigungu | 0.182 | 0.286 | 0.63 | 0.524 |

$\varphi = 0.065$, ***: $\alpha = 0.001$, **:$\alpha = 0.01$, *: $\alpha = 0.05$, §: $\alpha = 0.1$

### 3.3. 비교추정량

로지스틱 선형혼합모형을 이용한 EBLUP 타입 추정량의 성능은 다른 추정량들과 비교를 통해 평가되었다. 이때 비교 추정량은 직접추정량과 모형기반인 EBLUP과 HB(Hierarchical bayes) 추정량을 사용하였다. 본 연구가 사례연구이고, 비교가 핵심은 아니라는 점에서 추정량들의 성격을 간략하게 설명하기로 한다. 자세한 내용은 Rao (2003)를 참고할 수 있다.

#### (1) 직접추정량

직접추정량은 모집단의 보조정보인 층별 추계인구를 활용한 비추정(ratio estimation)을 사용한다. 비추정에 의한 $d$ 지역의 직접추정량 $\hat{y}_d^{dir}$는 다음과 같다.

$$\hat{y}_d^{dir} = \sum_j \sum_k \sum_l w_{djkl} y_{djkl},$$

여기서 $w_{djkl}$은 사후층화 가중값, $y_{djkl}$는 $d$ 지역의 $j$ 표본조사구내 $k$ 가구의 $l$번째 가구원의 관측값을 나타낸다. 직접추정량은 비편향추정량이지만, 소지역에 할당된 표본의 크기가 작은 경우에는 추정량의 분산이 커져서 신뢰성이 떨어지는 특성이 있다.

**(2) EBLUP 추정량**

EBLUP와 HB 추정량은 모형 추정량으로서, FH (Fay와 Herriot, 1979)모형을 사용한다. FH 모형은 일반선형혼합모형의 특수한 경우로, 다음과 같은 모형을 사용한다.

$$y_d = \mathbf{x}_d^t\beta + \nu_d + e_d = \theta_d + e_d, \quad e_d \sim N(0, \psi_d),\ \nu_d \sim N(0, \sigma_\nu^2), \tag{3.3}$$

여기서, $y_d$는 $d$ 지역에서 관측된 실업자수, $x_d$는 보조변수, $\nu_d$는 지역 랜덤 효과, $e_d$는 표집오차를 나타낸다. 그리고 $\theta_d = \mathbf{x}_d'\beta_d + \nu_d$는 $d$ 소지역의 참값으로 선형회귀모형을 따른다. 이때 $e_d$와 $\nu_d$는 서로 독립이다. $e_d$의 분산 $\psi_d$는 일반적으로 모르는 경우가 많기 때문에, 본 연구에서 $\psi_d$는 잭나이프 방법에 의해 자료로부터 추정하였다. 이때 반응변수 $y$는 실업자수로서 실제 분석에서는 자연로그로 변환한 값을 사용하였다.

지역랜덤효과 $\nu_d$의 분산 $\sigma_\nu^2$이 지역에 대해 동일하다고 가정하면, $\psi_d$와 $\sigma_\nu^2$에 대한 $\theta_d$의 EBLUP 추정량 $\hat{\theta}_d$은 식 (3.2)와 같다.

$$\hat{\theta}_d = \gamma_d\hat{y}_d + (1-\gamma_d)\mathbf{x}_d^t\hat{\boldsymbol{\beta}}, \tag{3.4}$$

$\hat{\boldsymbol{\beta}}$은 $\boldsymbol{\beta}$의 최소제곱추정치이고, 가중치는 $\gamma_d = \hat{\sigma}_\nu^2/(\psi_d + \hat{\sigma}_\nu^2)$이다. 여기서 $\hat{\sigma}_\nu^2$은 ML 추정치를 사용한다 (Rao, 2003). FH 모형에 의한 EBLUP 추정량은 가중치 $\gamma_d$를 계산할 때 표집오차, $\hat{\psi}_d$가 사용된다는 점에서 경활조사의 표본제외지역이나 표본조사구가 1개인 지역에서는 표집오차를 계산할 수 없다. 또한 EBLUP 추정량의 MSE(Mean Squared Error)는 과소 추정되는 경향이 있다는 연구결과가 이론적 또는 경험적으로 이미 알려져 있다 (Rao, 2003; 김서영과 권순필, 2009).

**(3) HB 추정량**

식 (3.3)을 소지역 추정을 위한 계층적 모형으로 표현하면 각각 다음과 같다. 즉, $\theta_d$에 대한 조건부 모형은

$$y_d|\theta_d \sim N(\theta_d, \sigma_d^2)$$

이고, $\beta,\ \sigma_\nu^2$에 대한 조건부모형은

$$\theta_d|\beta,\ \sigma_\nu^2 \sim N(x^t\beta, \sigma_\nu^2)$$

이다. 이때 $\beta$와 $\sigma_\nu^2$은 독립이고, $\sigma_\nu^2 \sim \mathrm{IG}(a, b)$, $\beta \propto 1$이다. 회귀계수 $\beta$에 대한 사전분포는 무정보적 사전분포(flat prior)를 사용한다. 또, 역감마분포는 $\sigma_\nu^2$의 공액사전분포이다 (Rao, 2003; 김달호, 2005). 본 연구에서는 $a = b = 0.001$을 사용하였다. 소지역 $d = 1, 2, \cdots, D$에 대해, 깁스 표본자를 사용하기 위한 조건부확률분포는 다음과 같다.

$$\beta|y,\ \sigma_\nu^2,\ \theta \sim N\left((X'X)^{-1}X'\theta,\ \sigma_\nu^2(X'X)^{-1}\right),$$

$$\theta_d|y,\ \beta,\ \sigma_\nu \sim N\left((1-r_d)y_d + r_d x_d'\beta,\ \sigma_d(1-r_d)\right), \quad r_d = \frac{\sigma_d^2}{\sigma_d^2+\sigma_\nu^2},$$

$$\sigma_\nu^2|y,\ \beta,\ \psi,\ u,\ \theta \sim \mathrm{IG}\left(a + \frac{D}{2},\ b + \sum_{d=1}^{D}\frac{(\theta_d - x_d'\beta)^2}{2}\right),$$

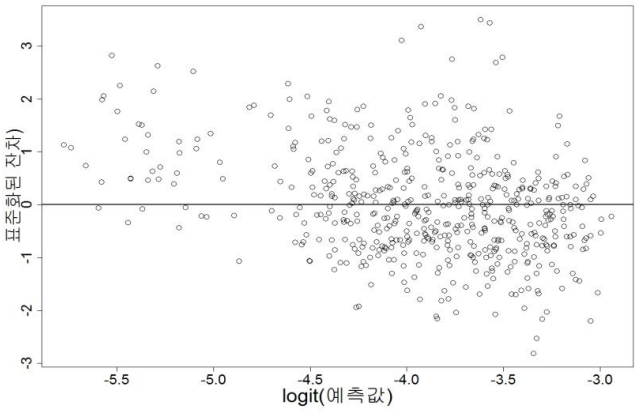



그림 **4.1.** 실업에 대한 잔차 플롯

여기서 관측치 벡터 $\boldsymbol{y} = (y_1, \ldots, y_D)'$, 보조변수 벡터 $X' = (x_1, x_2, \ldots, x_D)$, 모수벡터 $\theta' = (\theta_1, \theta_2, \ldots, \theta_D)'$를 나타낸다. 각 확률변수에 대한 조건부사후분포의 평균과 분산을 구하기 위해 깁스 샘플링을 이용한다 (Gelfand와 Smith, 1991). $d$ 소지역의 사후평균과 분산은 마크프체인 몬테칼로 방법에 의한 라오블랙웰 추정량(Rao-Blackwellized estimator)을 이용한다 (You 등, 2003). 이에 대한 자세한 내용은 You 등 (2003), 김달호 (2005)를 참고할 수 있다.

본 연구에서 사용한 로지스틱 선형혼합모형과 FH 모형은 약간 다른 형태를 갖는다. 두 모형 모두 지역기반 모형이라는 점은 같지만, 추정 단위에 있어서 FH 모형은 해당 소지역 $d$의 실업자수, 로지스틱 모형은 소지역 $d$의 성×연령별 그룹의 실업자수라를 추정한다는 점에서 차이가 있다. 따라서 로지스틱 모형 기반 EBLUP 타입 추정량이 다른 추정량들과의 비교에서 절대적 우월성은 평가하기 어려울 것이다.

## 4. 결과

모형진단과 추정량들의 평가를 위해 다양한 척도를 사용하였다. 추정량 평가는 반드시 이론적인 기반을 두지 않더라도 다양한 시각적 기법을 이용해서 추정량의 특성 등을 살펴볼 수 있다 (Brown 등, 2001; Heady 등, 2003).

### 4.1. EBLUP 타입 추정량의 절대적 평가

#### (1) 모형진단

그림 4.1은 예측된 실업자수에 대해 R에서 제공한 표준화된 잔차(standardized residual)를 그린 것이다. 그림에서 보면 잔차들의 분포가 0을 중심으로 ±3 사이에 고루 퍼져있어 모형추정에 의해 예측치의 특이한 현상은 없다고 볼 수 있다.

#### (2) 편향 측정

모형기반 소지역 추정치의 편향은 직접추정치와 모형추정치와의 선형성을 통해 측정할 수 있다. 회귀식, $y$(직접추정치) $= \beta_0 + \beta_1 x$(모형추정치)에 대해 계수 $\beta = 0$, $\beta_1 = 1$을 검정하고, 회귀선이 $y = x$에 일치할수록 추정치의 편향은 없다고 판단한다. 그림 4.2는 모형에 의한 EBLUP 타입 추정치와 직접추정치와 선형관계를 나타낸 것이다. 각 추정치는 로그변환 되었다. 그림에서 실선은 $y = x$ 선이고, 점선

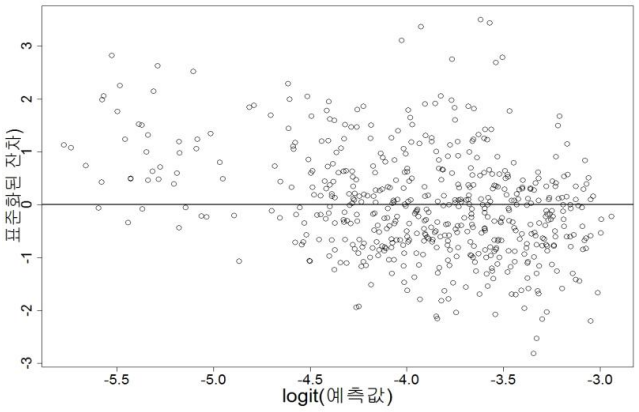



**그림 4.2.** 모형추정치와 직접추정치의 선형성

은 추정된 $y = -0.12(0.28) + 1.01(0.03)x$와 같다. 이때 괄호는 추정된 회귀계수에 대한 표준편차를 나타낸다. 각 회귀계수 결과, 유의수준 $\alpha = 0.05$에서 위의 가설을 모두 채택하였다. 따라서 추정된 회귀선은 $y = x$와 거의 일치하고, EBLUP 타입 추정치는 직접추정치에 대해 크게 편향되지 않는다고 볼 수 있다.

### (3) 상대표준오차 계산

상대표준오차, CV는 추정치의 신뢰성을 측정하는 지표로 사용된다. 일반적으로 소지역 통계를 공표할 경우, 상대표준오차는 20% 또는 25% 기준이 자주 사용된다. 본 연구에서는 상대표준오차 25% 기준을 사용하였다. $d$ 지역에 대한 상대표준오차는 다음과 같이 계산된다.

$$\mathrm{CV}_d = \frac{\sqrt{\mathrm{MSE}_d}}{\hat{\theta}_d} \times 100, \quad \hat{\theta}_d : d\text{지역의 소지역 추정치}.$$

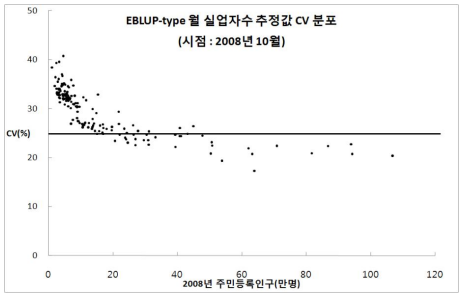
그림 4.3은 월, 분기, 연자료에 대한 EBLUP 타입 추정치의 상대표준오차값을 그린 것이다. 월 추정치의 경우, 특광역시내의 구 지역을 제외한 많은 시군 지역에서 상대표준오차가 25%보다 큰 것으로 나타났다. 특히 군 단위의 농어촌 지역 또는 인구밀도가 낮거나 인구가 적은 지역은 상대표준오차가 상당히 큰 것으로 나타났다. 실제로 경활조사 대부분 소지역은 표본수가 매우 작기 때문에 매월 조사되는 조사치 자체의 변동이 클 수 있고, 이런 측면에서 보면 월 추정치의 상대표준오차가 큰 것이 그렇게 이상한 것은 아닌 것 같다. 한편, 그림 4.3(b), (c)의 분기 또는 연 추정치는 월 추정치에 비해 상대표준오차가 상대적으로 작아지는 것을 알 수 있다.

## 4.2. EBLUP 타입 추정량의 상대성 평가

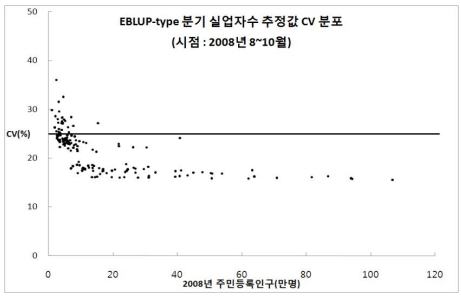
본 절에서는 EBLUP 타입 추정치와 다른 추정치들을 상대적으로 비교하였다. 비교를 위해 대표본 조사추정치, 직접추정치, EBLUP 추정치, 또는 HB 추정치를 사용하였다. 대표본 조사추정치는 지역별 고용조사에 의한 조사추정치를 말한다.

### (1) 추정치 분포

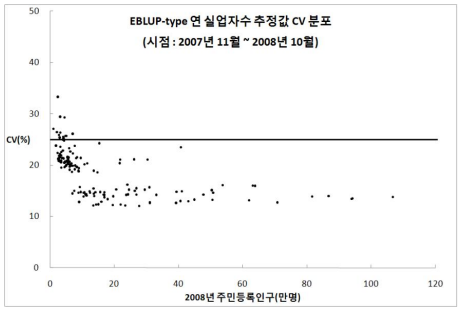
그림 4.4는 지역별 고용조사 추정치, 로지스틱 모형의 EBLUP 타입 추정치, FH 모형의 EBLUP 추정치를 나타낸 것이다. 그림 4.4에서 $x$축은 표본조사구수, $y$축은 실업자수를 나타낸다. 그림을 보면, 각




(a) 월 자료




(b) 분기 자료




(c) 연 자료

**그림 4.3.** 월, 분기, 연 자료의 추정치에 대한 상대표준오차 값

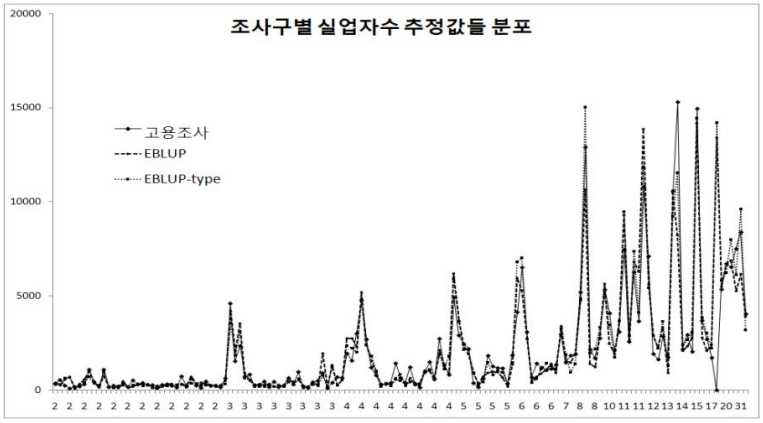



**그림 4.4.** 각 추정치들의 분포

추정치들은 표본조사구수에 따라 전체적인 경향이 비슷하다는 것을 알 수 있다. 특히 EBLUP 타입 추정치는 지역별 고용조사(고용조사)추정치에 대해 근사한 것으로 나타났다. 표본조사구수가 증가함에 따라 EBLUP 타입 추정치가 다른 추정치들에 비해 크게 추정되는 지역도 있는데, 이 지역들 중에는 표본조사구수가 8개, 5개, 6개로 대체로 표본이 큰 지역들도 포함되어 있다. 따라서 이런 경우는 지역특성을 반영할 수 있는 보조변수를 추가하여 추정치의 신뢰성을 높이거나, 소지역 추정치를 해석할 때 지역특성을 잘 고려할 필요가 있을 것이다.

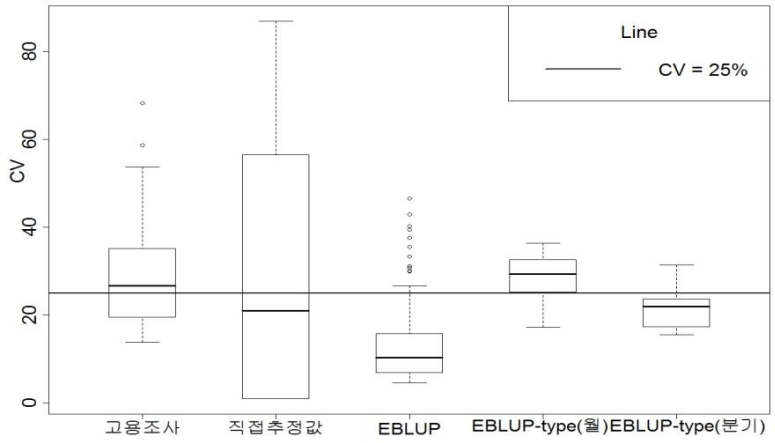



**그림 4.5.** 각 추정치들의 분포

### (2) 상대표준오차 분포

그림 4.5는 세 추정량의 상대표준오차 분포를 나타낸다. EBLUP 타입 추정치에 대해서는 월과 분기 추정치의 상대표준오차를 모두 비교하였다. 우선 모든 모형 추정치들은 직접추정치의 상대표준오차를 대폭 개선하는 것을 알 수 있다. 또한 EBLUP 타입 추정치와 EBLUP 추정치의 상대표준오차는 지역별 고용조사의 상대표준오차보다 더 작게 나타났다. EBLUP 타입 추정치의 경우, 월 추정치보다는 분기 추정치의 상대표준오차가 상대적으로 더 작다. 그림 4.5의 결과에서는 EBLUP 추정치의 상대표준오차가 가장 작게 나타났지만, EBLUP 추정치의 MSE가 과소 추정될 수 있다는 점에 주의할 필요가 있겠다. 이런 측면에서 볼 때, EBLUP 타입 추정치는 상대표준오차 25% 기준에서 대표본조사보다 더 작고, 상자의 폭이 좁은 것으로 보아 지역에 따른 상대표준오차의 변동성도 상대적으로 작다고 볼 수 있다.

### (3) 대지역 추정치 비교

신뢰가능한 대지역 단위에서의 직접추정치와 소지역 추정치간의 비교함으로써 소지역 추정치의 신뢰성을 확인할 수도 있다. 비교를 위해 경활조사의 시도 단위 직접추정치와 소지역 추정치를 시도 단위로 집계한 값을 사용한다.

그림 4.6은 각 추정치의 95% 신뢰구간을 각 시도별로 그린 것이다. 그림 4.6에서 95% 신뢰구간은 경활조사의 상대표준오차를 알려진 값으로 간주하여 계산하였다. 그림에서 보는 바와 같이 대부분 시도의 EBLUP 타입 추정치의 신뢰구간은 경활조사 신뢰구간과 유사하고 두 구간이 거의 겹치는 것을 알 수 있다. 경기도와 서울의 경우 두 신뢰구간이 약간 어긋나 있는데, 이는 소지역 추정치의 신뢰구간을 계산할 때 직접추정치의 상대표준오차를 추정치로 사용했다는 것도 하나의 이유가 될 수 있다. 이 그림을 통해서 상대표준오차 측면에서 EBLUP 타입 소지역 추정치는 신뢰할 만하다고 볼 수 있다.

### (4) 시계열 확인

연속조사의 경우 시계열은 자료의 변동성을 확인하고 그로부터 변동성의 이유를 파악하는데 중요가 근거가 될 수 있다. 소지역 추정치를 시계열 측면에서 평가하였다. 이때 2005년부터 2008년까지의 3년간의 분기자료를 사용하였다. 보조변수로 활용되는 행정자료 사용의 제한으로 더 긴 시계열을 추정할 수

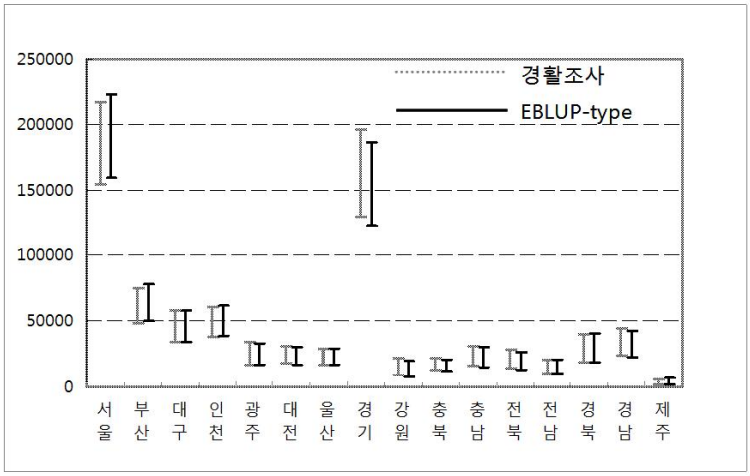



**그림 4.6.** 시도단위에서 EBLUP 타입 추정과 경활조사 실업자수에 대한 신뢰구간

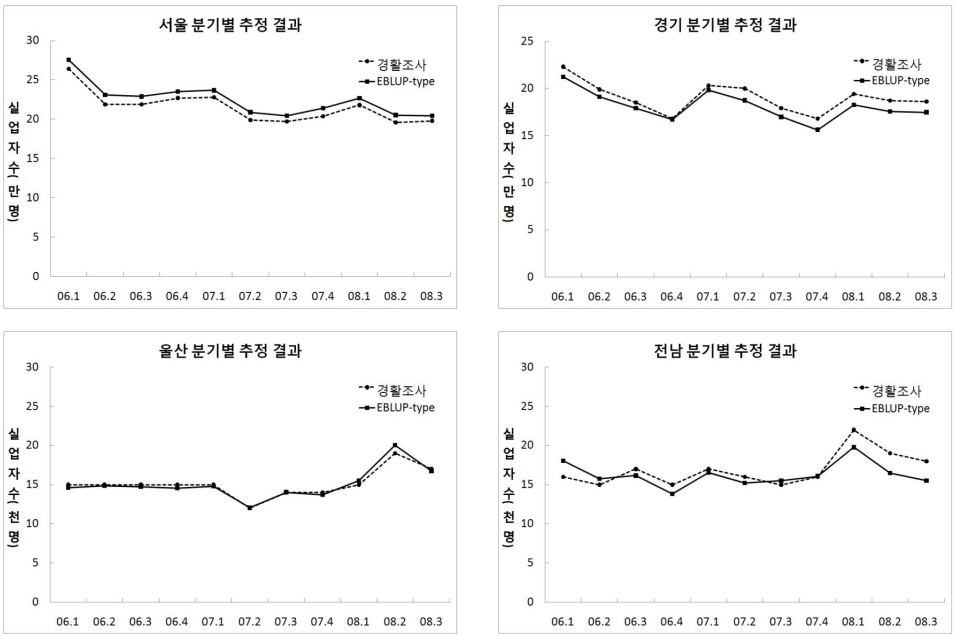



**그림 4.7.** 경제활동인구조사와 EBLUP 타입 추정치의 분기 시계열 도표

는 없었다. 추정을 통해 16개 모든 시도의 경활조사 실업자수와 EBLUP 타입 실업자수 추정치의 시계열을 비교하였다. 다만, 지면의 한계 상 본 논문에서는 서울, 경기, 울산, 전남 지역 4개 지역을 임의로 선정하여 그 결과를 제시하였지만 다른 지역들에서도 결과는 유사하였다.

그림 4.7은 4개 지역들에 대한 경활조사 직접추정량과 EBLUP 타입 추정량에 의해 추정된 실업자수의 시계열을 나타낸 것이다. 전체적으로 각 지역들에서 경활조사와 EBLUP 타입 추정치의 시계열이 유사

표 **4.1.** 시스템 통제의 적정성 판단을 위한 질문항목

| 지역 | ARB | | | RMSE(%) | | |
|---|---|---|---|---|---|---|
| | EBLUP 타입 | EBLUP | HB | EBLUP 타입 | EBLUP | HB |
| 1 | 0.37 | 0.47 | 0.31 | 48.32 | 51.84 | 38.82 |
| 2 | 0.40 | 0.51 | 0.29 | 59.65 | 66.00 | 36.28 |
| 3 | 0.51 | 0.85 | 0.43 | 67.95 | 94.86 | 50.46 |
| 4 | 0.62 | 0.88 | 0.49 | 76.24 | 96.67 | 56.46 |
| 5 | 0.42 | 0.81 | 0.38 | 62.84 | 98.56 | 45.53 |
| 6 | 0.57 | 0.80 | 0.47 | 73.52 | 97.18 | 59.66 |
| 7 | 0.71 | 0.90 | 0.46 | 97.79 | 104.56 | 55.68 |
| 8 | 0.43 | 0.64 | 0.36 | 61.98 | 70.75 | 43.70 |
| 9 | 0.23 | 0.58 | 0.21 | 29.10 | 61.82 | 25.92 |
| 평균 | 0.44 | 0.65 | 0.37 | 58.88 | 71.12 | 42.18 |

한 것을 알 수 있다. 지역적으로 보면, 전남은 2008년에 접어들면서 경활조사와 소지역 추정치간의 차이가 약간 발생하는 것을 알 수 있다. 서울은 모든 분기에서 소지역 추정치가 경활조사 추정치보다 약간 높게 추정되었고, 경기는 소지역 추정치가 경활조사 추정치보다 전체적으로 약간 낮게 추정되었다. 이에 대한 한 가지 이유로 서울과 경기도는 인접 지역으로 동일한 출퇴근 권역에 있다는 점을 들 수 있다. 이처럼 모형추정에 있어서 지리적 경계선상에 있는 지역들은 그 특성이 명백하게 구분되지는 않기 때문으로, 추정하는데 상당히 어려움이 따를 수 있다. 따라서 모형을 이용하여 지역 특성을 추정할 경우, 인접 지역들의 특성을 모형에 반영할 수 있는 방법을 찾는 것도 중요한 과제일 것이다.

### 4.3. 모의실험에 의한 추정치 평가

본 절에서는 EBLUP 타입 추정량의 불편성을 측정하기 위해 모의실험을 수행하였다. 2005년 센서스 자료를 모집단으로 간주하고, 이로부터 1,000개의 표본을 복원 추출하여 실업자수를 추정하고 이를 센서스 자료의 실업자수와 비교하였다. 이때 매번 실험에서의 표본은 경활조사의 표본추출과 동일한 방법으로 동일한 크기의 표본을 추출하였다. EBLUP 타입 추정량의 성능은 통상적으로 사용되는 척도인 평균 절대 상대편향(average Absolute Relative Bias; ARB)과 상대 평균제곱오차(Relative Mean Squared Error; RMSE)를 이용하여 평가하였다.

$y_d$는 소지역 $d$의 모수, $\hat{y}_{d,r}$은 $r$번째 표본에서 계산된 소지역 $d$의 추정치라 할 때, ARB와 평균 ARB는 다음과 같다.

$$\mathrm{ARB}_d = \frac{1}{R}\left(\sum_{r=1}^{R}\frac{|\hat{y}_{d,r}-y_d|}{y_d}\right), \qquad \overline{\mathrm{ARB}} = \frac{1}{D}\sum_{d=1}^{D}\mathrm{ARB}_D, \quad D\text{는 소지역 개수},$$

ARB는 0에 가까울수록 좋고, 상대적인 편향이 작음을 의미한다. 소지역 $d$에 대한 RMSE와 평균 RMSE는 다음과 같다.

$$\mathrm{RMSE}_d = \frac{\sqrt{\mathrm{MSE}_d}}{y_d}\times 100, \qquad \mathrm{MSE}_d = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{y}_{d,r}-y_d\right)^2, \qquad \overline{\mathrm{RMSE}} = \frac{1}{D}\sum_{i=1}^{d}\mathrm{RMSE},$$

RMSE는 0에 가까울수록 좋고, 표본에 대한 추정량들의 산포도를 측정할 수 있다.

표 4.1은 EBLUP, EBLUP 타입, HB 추정량에 대해 특광역시를 제외한 9개 시도와 전국 ARB, $\overline{\mathrm{ARB}}$, RMSE, $\overline{\mathrm{RMSE}}$를 나타낸다. 표에서 보면, 세 추정량 중 HB의 편향이 가장 작고, 그 다음이 EBLUP 타

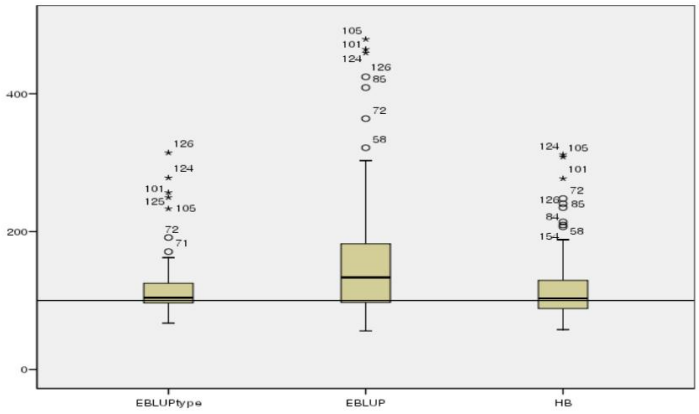



**그림 4.8.** 세 추정량들의 편향 분포

입, EBLUP 추정량 순으로 나타났다. 효율성은 HB의 $\overline{\text{RMSE}}$가 가장 낮고, 그 다음으로 EBLUP 타입, EBLUP 순으로 낮게 나타났다. 따라서 편향성과 효율성 면에서 HB가 가장 좋고, EBLUP 타입이 그 뒤를 이었다. HB 추정량은 사전분포가 잘 정의된다면 그 추정량의 효율성이 좋다는 것은 이미 이론적으로 많은 선행연구들에서 증명되었다 (You 등, 2003 등). 단지, HB 추정량은 추정 과정에서 마코프체인 몬테칼로 시뮬레이션이 사용된다는 점에서 국가통계 사용에 신중해야 한다는 의견들도 적지 않은 것 같다.

표 4.1을 보면, 편향은 HB가 0.37로 EBLUP 타입 추정량의 0.44에 비해 약간 좋은 편이다. 한편 편향의 상대적 분포는 HB가 EBLUP 타입에 비해 변동성이 더 큰 것으로 나타났다.

그림 4.8은 전 지역에 걸쳐 수행한 1,000번의 실험결과에 대해 EBLUP 타입, EBLUP, HB 추정량들의 편향에 대한 상대적 변동성을 나타낸 것이다. 그림에서 변동성이 큰 순서로 보면, EBLUP, HB, EBLUP 타입 순인 것을 알 수 있다. 평균적인 개념에서는 HB의 편향이 더 작았지만(표 4.1), 모의실험에서 지역 전체를 놓고 보면 EBLUP 타입 추정량이 HB보다 비교적 안정적인 것을 알 수 있다. 지역적 특성이나 추출된 표본에 따라 HB 추정량은 EBLUP 타입 추정량에 비해 좀더 영향을 받기 쉽다고 볼 수 있다. 모의실험 결과에서 EBLUP 타입 추정량은 HB 또는 EBLUP 추정량에 대해 편향성 측면에서는 성능이 비슷하거나 좋고, 실용성과 안정성 측면에서 더 우수하다고 볼 수 있다.

## 5. 결론 및 논의

지금까지 로지스틱 선형혼합모형 기반 EBLUP 타입 추정량을 이용하여 우리나라 시군구 실업자수를 추정해 보았다. 경제활동인구조사의 실업자수 자료를 종속변수로 하고, 성, 연령, 지역 가변수와 행정자료인 실업급여등록자수를 보조변수로 사용하였다. 추정 결과에 따르면 로지스틱 선형혼합모형 기반 EBLUP 타입 추정량은 모형적합성, 편향성, 신뢰성, 조사추정치와의 유사성 측면에서 볼 때 그 활용 가능성이 충분히 확인되었다고 볼 수 있다. EBLUP 타입 추정량은 EBLUP 추정량 또는 HB 추정량에 비해 그 성능이 우수하거나 최소한 비슷한 것으로 평가되었다. 시점별 추정의 경우 경활조사를 이용한 소지역 실업자수 추정은 상대표준오차의 안정성 측면에서 볼 때 월 추정치보다는 분기 추정치가 훨씬 추천될만하다.

본 연구에서 사용한 방법 하에서 소지역 추정치의 신뢰성을 향상시킬 수 있는 방법을 두 가지 정도 생

각해 볼 수 있겠다. 먼저 가능하다면 시군구에서 최소한의 표본크기를 확보하는 것이다. 소지역 추정의 경우, 소지역에서의 표본크기가 너무 작으면 추정치에 대한 신뢰성이 감소할 수 있기 때문에, 경활조사 표본설계시 소지역 추정을 전제로 소지역의 표본을 할당한다면 적은 비용으로 추정치의 정도를 향상시킬 수 있다. 다른 하나는 소지역 정의 문제이다. 본 연구에서와 같이 행정구역 단위를 그대로 사용하지 않고, 좀 더 큰 권역의 형태로 묶어서 추정할 수 있을 것이다. 이때 어떻게 권역을 설정할 것인가 하는 것도 고민해야 할 문제이다. 이 경우, 통합그룹 안에 포함된 각 소지역들에 대해서는 추정치의 정도를 계산할 수 없다는 제약이 따를 수 있다. 마지막으로 보조변수 선택이 소지역 추정의 중요한 관건이 되고 있다. 본 연구에서는 소지역의 실업급여등록자라는 행정자료를 보조변수로 사용하였다. 본문에서는 언급하지는 않았지만, 센서스의 실업자수라든가 소지역을 포함한 상위 지역의 실업자수를 보조변수로 사용하였을 때 추정치는 실제로 크게 향상시키지는 않았다. 하지만 다른 행정자료나 지리정보 개념의 보조변수를 찾는 연구는 꾸준히 지속될 필요가 있다.

소지역 추정치의 포괄범위에 있어서 전국 단위 소지역 추정은 해당 소지역의 특성을 모형에 잘 반영할 수 있을 때, 그 추정치의 신뢰성은 훨씬 향상될 수 있다. 추가분석에 따르면 소지역 추정치가 대표본 조사의 95% 신뢰구간에 포함되지 않은 지역들은 의정부시, 하남시, 통영시, 밀양시, 화순군, 동해시, 삼척시 등 해안지역이거나 관광지역적 특성이 포함된 지역들이라 할 수 있다. 해안도시의 경우 어업이나 관광지역이 많고, 이런 지역들의 산업적 특성은 다른 지역들에 비해 계절적 영향을 받기 쉬울 수 있다. 따라서 같은 소지역에 대해서 표본이 많은 지역통계 결과와는 다소 차이가 있을 수 있다. 이에 대해 사용자들이 이런 지역들을 해석할 때 주의하도록 안내하는 것도 좋은 활용방법이 될 것이다.

학문적 또는 실용적 측면에서 소지역 추정은 상당한 매력을 갖는다. 특히 국가통계로서의 매력은 첫째, 비용절감에 있다. 소지역 추정은 적은 비용으로 소지역 통계를 만들어 낼 수 있는 방법이다. 물론 조사통계만큼 다양한 정보를 얻을 수는 없겠지만, 필요한 목적에 한해서는 큰 비용 없이도 신뢰할만한 소지역 통계를 만들 수 있다는 장점이 있다. 둘째, 비표본오차에 대한 고민을 줄일 수 있다. 조사통계에서 해결하기 어려운 문제 중 하나는 비표본오차를 어떻게 측정하고 이를 어떻게 줄일 것인가 하는 것이다. 특히, 조사단위가 세분화되고 작아질수록 많아지는 표본과 함께 조사원 규모가 커지면서 조사는 그만큼 어려워지기 마련이다. 이럴수록 비표본오차가 언제 어떻게 얼마나 발생하는지를 예측하기는 매우 어려운 일일 것이다. 셋째, 시의성이다. 모든 통계는 제때에 이용자에게 전달되었을 때 그 가치가 돋보일 것이다. 소지역 추정 기법은 자료만 주어진다면 추정결과를 빠르게 제공할 수 있는 방법이다.

본 연구와 관련하여 소지역 추정치의 신뢰성을 향상시키기 위해 향후 꾸준한 연구가 진행되어야 할 것이다. 우선 대지역 단위에서 경활조사 추정치와 소지역 추정치를 일치시킬 수 있는 벤치마킹방법과 추정치의 정도 향상을 위한 RMSE를 개선 방법을 찾는 연구를 수행할 예정이다. 이처럼 소지역 추정이 하나의 통계작성 방법으로 자리매김하기까지는 이 분야의 많은 연구자들의 꾸준한 노력과 함께 소지역 추정에 대한 사용자의 인식 전환이 중요한 요소가 될 것으로 보인다.

## 참고문헌

김달호 (2005). <R과 Winbugs를 이용한 베이지안 통계분석>, 자유아카데미.

김서영, 권순필 (2009). <고용통계 소지역 추정 연구 Ⅰ>, 통계개발원 보고서.

김서영, 권순필 (2010). Time series & cross sectional 모형 기반 소지역 추정 사례연구, <통계연구>, 28–43.

김수택, 고석남, 김상대 (2008). 소지역 노동통계의 효율적 추정방안- 실업률을 중심으로-, <산업관계연구>, **18**, 53–76.

김정숙, 황희진, 신기일 (2008). 이웃정보시스템을 이용한 공간 소지역 추정량 비교, <응용통계연구>, **21**, 855–866.

박종태, 이상은 (2001). 소지역 추정법에 관한 비교 연구, <한국데이터정보고학회지>, **12**, 47–55.

이계오 (2000). 시군구 실업자 추정을 위한 소지역 추정법, <응용통계연구>, **13**, 276–286.

이상은 (2006). 공간통계량을 활용한 베이지안 자기포아송 모형을 이용한 소지역 통계, <응용통계연구>, **19**, 424–430.

정연수, 이계오, 이우일 (2003). 시군구 실업자 통계추정을 위한 설계기반 간접추정법, <응용통계연구>, **16**, 1–14.

통계청 (2008). http://kosis.kr/nsp/index/index.jsp.

Brown, G., Chambers, R., Heady, P. and Heasam, H. (2001). Evaluation of small area estimation methods-an application to the unemployment estimates from the UK LFS. In *Proceedings of Statistics Canada Symposium Ottawa, October 2001*.

Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data, *Journal of the American Statistical Association*, **74**, 269–277.

Gelfand, A. E. and Smith, A. F. M. (1991). Gibbs sampling for marginal posterior expectations, *Communications in Statistics-Theory and Methods*, **20**, 1747–1766.

Heady, P., Clarke, P., Brown, G., Eillis, K., Heasman, D., Hennell, S., Longhurst, J. and Mitchell, B. (2003). Model-based small area estimation series NO.2: Small area estimation project report, Office for National Statistics (UK) publication.

KOSIS (2010). http://kosis.kr/metadata/.

Molina, I., Saei, A. and Lombardia, M. J. (2007). Samll area estimates of labour force participation under a multinomial logit mixed model, *Journal of the Royal Statistical Society: Series A*, **170**, 975–1000.

ONS (2009). Model-Based Estimates of ILO Unemployment for UA/LADs in great Britain Guide for Users.

Rao, J. N. K (2003). *Small Area Estimation*, Wiley.

Rao, J. N. K. and Yu, M. (1994). Small area estimation by combining time series and cross-sectional data, *The Canadian Journal of Statistics*, **22**, 511–528.

Saei, A. and Chambers, R. (2003a). EBLUP-type estimate of small area parameters based on logistic models, Working paper of University of Southampton, July 2003.

Saei, A. and Chambers, R. (2003b). Small area estimation: A review of methods based on the application of mixed models, Working paper of University of Southampton, September 2003.

Saei, A. and McGilchrist, C. (1998). Longitudinal threshold models with random components, *The Statistician (JRSS Series D)*, **47**, 365–375.

Tiller, R. B. (1992). Time sereies modeling of sample survey data from the U.S.Current Population Survey, *Journal of Official Statistics*, **8**, 149–166.

You, Y., Rao, J. N. K. and Gambino, J. (2003). Model-based unemployment rate estimation for the canadian labour force survey: A Hierarchical bayes approach, *Statistics Canada*, **29**, 25–32.

# Evaluation of EBLUP-Type Estimator Based on a Logistic Linear Mixed Model for Small Area Unemployment

Seo Young Kim[1] · Soon Pil Kwon[2]

[1]Statistics Research Institute, Statistics Korea; [2]Statistics Research Institute, Statistics Korea

(Received April 2010; accepted August 2010)

**Abstract**

In Korea, the small area estimation method is currently unpopular in generating official statistics. Because it may be difficult to determine the reliability for small area estimation, although small area estimation has a sufficiently good advantage to generate small area statistics for Korea. This paper inspects the method of making small area unemployment through the small area estimation method. To estimate small area unemployment we used an EBLUP-type estimator based on a logistic linear mixed model. To evaluate the EBLUP-type estimator we accomplished the real data analysis and simulation experiment from the population and housing census data. In addition, small area estimates are compared to large sample survey estimates. We found the provided method in this paper is highly recommendable to generate small area unemployment as the official statistics.

Keywords: Small area estimation, unemployment, generalized linear mixed model.

[1]Corresponding author: Ph.D, Statistics Research Institute, Daejeon Narakeyum 282-1 Wolpyeong-dong, Seo-gu, Daejeon 302-280, Korea. E-mail: smilegong@korea.kr.