한 국 통 계 학 회 논 문 집
2010, 17권, 3호, 337–347

시군 실업통계 작성을 위한 소지역 추정모형

김 진[1,a], 김재광[b]

[a]통계청 표본과, [b]아이오와주립대학교 통계학과

요 약

대부분의 통계조사는 전국 및 시도단위와 같은 전체 또는 규모가 큰 지역을 위한 통계작성이 가능하도록 설계되어 있다. 그러나 최근들어 소규모 지역의 신뢰있는 통계결과에 대한 통계 이용자들의 요구가 점점 증가하고 있다. 경제활동인구조사는 전국 및 16개 시도를 위한 실업률을 매월 작성하고 있으나, 시군별 직접추정량은 적은 표본규모 때문에 신뢰도가 낮다. 본 연구에서는 경제활동인구조사에서 제공되는 실업자수에 대한 통계를 시군별로 작성하기 위한 모형을 제시한다. 이를 위해 보조정보로부터 "정보를 빌려오는(borrow strength)" 지역단위모형(Area level model)을 사용하였다. 제안된 모형 기반 추정(Model-based Estimation)은 모형 오차의 정규성 가정을 기반으로 하고 있으며, 평균제곱오차 측면에서 제안된 모형들을 비교하였다.

주요용어: 지역단위 모형, 간접추정, 모형기반추정, 평균제곱오차.

1. 서론

표본조사에서는 주요 공표범위에 대한 통계작성을 위해 주어진 허용오차를 갖는 최소의 표본규모를 산출한다. 많은 통계이용자들은 좀 더 세분화된 통계결과를 이용하고자 하지만 표본규모는 비용과 직결되므로 통계 생산자 입장에서는 비용을 무한정 늘려 표본규모를 확대한다는 것은 매우 어려운 일이다. 이러한 이유로 우리나라의 대부분의 승인통계는 전국 또는 시도별 통계결과를 작성하고 있다. 또한 전국단위 통계에 대해서는 각 시도자료의 요구, 시도단위 통계에 대해서는 시군구 및 시도내 세분화된 자료의 요구가 점차적으로 증가하고 있는 실정이다.

경제활동인구조사는 우리나라의 취업자수 및 실업자수 등 고용통계를 전국 및 16개 시도별로 생산하고 있다. 그러나 IMF를 겪으면서 1998년 이후 고용문제가 사회의 큰 이슈로 자리잡게 되었고, 지자체에서는 해당 시도 뿐 아니라 시군구에 대한 고용통계 작성을 지속적으로 요구하여 왔다. 이에 통계청에서는 소지역추정법에 의한 시군구 실업통계 개발 (통계청, 2002) 및 모형기반 소지역추정 (통계청, 2004) 등 경제활동인구조사 결과자료를 이용하여 복합추정법(Composit Estimation)과 모형기반 추정법(Model-based Estimation)을 적용한 시군구 실업자수 추정 가능성을 시도한 바 있다.

국내에서도 경제활동인구조사 자료를 이용한 소지역추정에 관한 많은 연구가 이루어졌다. 이계오 (2000), 정연수 등 (2003), Chung 등 (2003) 는 충북지역의 시군별 실업자수를 복합추정식에 의해 추정하고 잭나이프방법에 의한 분산추정식을 제시하였고, 이상은 (2001, 2006)은 소지역추정법을 위한 베이즈방법과 공간통계량을 활용한 방법을 제안하였다.

해외의 연구동향을 살펴보면, 캐나다, 미국, 유럽 등 많은 국가에서 소지역추정에 대한 연구를 지속적으로 진행하고 있다. 우리나라에서 소지역추정이 경제활동인구조사 자료에 지역단위모형을 적용한 시군구 실업자수를 추정하는 단계라고 한다면, 많은 통계선진국에서는 10여년 전에 그 단계를 마치고

[1] 교신저자: (302-701) 대전시 서구 선사로 139 정부대전청사 3동, 통계청 표본과, 사무관. E-mail: jinknso@korea.kr

소지역추정에서의 오차측정 (Gonzalez-Manteiga 등, 2006; Wang과 Fuller, 2003; Ybarra와 Lohr, 2008), 범주형 자료에서의 소지역추정기법 (Hall과 Maiti, 2006; Opsomer 등, 2008), GIS를 이용한 소지역추정기법 (Battese 등, 1998) 등 다양한 연구분야로 확대되고 있다.

본 연구에서는 2005년 11월의 경제활동인구조사 자료를 이용하여 센서스 및 인구특성에 관한 행정자료를 보조변수로 활용한 모형기반 소지역추정기법인 경험적 베이즈방법(EB; Empirical Bayes)을 적용하여 실업자수를 추정하였다. 또한 모형의 정규성 가정을 만족하기 위하여 실업자수에 대한 로지스틱 변환을 실시하였으며 평균제곱오차(MSE)에 의해 각 결과들일 비교하였다.

2. 소지역모형 개요

소지역모형은 조사단위의 개별값으로 모형화한 단위수준모형(Unit level model)과 지역단위의 값으로 모형화한 지역수준모형(Area level model)으로 분류할 수 있다.

단위수준 모형(Unit level model)은

$$y_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + v_i + e_{ij}, \quad j = 1, \ldots, N_i, \;\; i = 1, 2, \ldots, m$$

으로, 조사단위의 자료값과 이와 관련된 보조자료를 결합한 모형이다. 이 모형은 조사단위별로 자료값과 보조자료가 존재한다.

지역수준 모형(Area level model)은

$$\hat{\theta}_i = \boldsymbol{z}_i{}^T\boldsymbol{\beta} + v_i + e_i, \quad i = 1, 2, \ldots, m, \;\; \boldsymbol{z}_i = (z_{1i}, \ldots, z_{pi})^T$$

으로 지역과 관련된 직접추정량과 보조자료를 결합한 모형이다. 여기서, m은 소지역 수, p는 보조정보의 수이다. 이 모형은 단위수준자료(Unit level data)를 얻을 수 없을 때 사용하며, 가구소득, 고용통계 등 대부분의 소지역모형이 지역수준 모형에 해당된다.

2.1. 지역수준 모형(Area level model)의 기본모형

Fay-Harriot 모형이라고 불리우는 지역수준 모형의 기본모형은 (2.1) 직접추정량과 모수의 모형과 (2.2) 모수와 보조변수간의 선형회귀모형을 결합한 형태인 (2.3)의 모형이다.

$$\hat{\theta}_i = \theta_i + e_i, \quad i = 1, 2, \ldots, m \tag{2.1}$$

식 (2.1)은 직접추정량 $\hat{\theta} = g(\hat{\bar{Y}}_i)$과 모수 $\theta_i = g(\bar{Y}_i)$의 모형이다. $\boldsymbol{e} = [e_1, \ldots, e_m]^T$는 표본오차로 평균 0, 분산 $V = [V_1, \ldots, V_m]^T$으로 표본오차 벡터이며, V_i는 알려진 것으로 가정한다.

$$\begin{aligned} \theta_i &= z_{i1}\beta_1 + z_{i2}\beta_2 + \cdots + z_{ip}\beta_p + v_i \\ &= \boldsymbol{z}_i^T\boldsymbol{\beta} + v_i, \quad i = 1, 2, \ldots, m \end{aligned} \tag{2.2}$$

식 (2.2)는 모수 θ_i와 지역단위 보조변수 $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \ldots, z_{ip})^T$의 선형회귀모형으로 지역단위모형에서 연결모형의 역할을 한다. v_i는 모형오차로 평균 0, 분산 $\sigma_v{}^2$를 갖는 *i.i.d.*이며, 표본오차 e_i와는 서로 독립임을 가정한다.

$$\hat{\theta}_i = \boldsymbol{z}_i^T\boldsymbol{\beta} + v_i + e_i, \quad i = 1, 2, \ldots, m \tag{2.3}$$

식 (2.3)은 지역수준의 소지역모형으로 Fay-Harriot 모형이라 하며, 식 (2.1)과 (2.2)의 결합모형이다 (Rao와 Yu, 1994; Rao, 2003).

지역수준모형의 여러 가지 확장모형(Expanded Area level Model)으로 다변량 Fay-Harriot 모형(Multivariate Fay-Harriot Model), 상관된 표본오차 모형(Model for correlated sampling error), 시계열 및 횡단면모형(Time series and cross-sectional Models)이 있다.

다변량 Fay-Harriot 모형은 $\hat{\boldsymbol{\theta}}_i = \boldsymbol{z}_i{}^T\boldsymbol{\beta} + \boldsymbol{v}_i + \boldsymbol{e}_i$으로, $\hat{\boldsymbol{\theta}}_{\boldsymbol{i}} = (\hat{\theta}_{i1}, \ldots, \hat{\theta}_{ir})^T$이 행렬형태이고, $\boldsymbol{v}_i$, $\boldsymbol{e}_i$도 행렬이다. 상관된 표본오차 모형은 $\hat{\theta} = \boldsymbol{z}^T\boldsymbol{\beta} + v + e$으로 $e|\theta \sim N_m(0, \boldsymbol{V})$, $\boldsymbol{V} = (V_{ij})$는 표본오차에 대한 $m \times p$인 공분산행렬을 나타낸다. 시계열 및 횡단면모형(TSCS 모형)은 연결모형이 시계열모형인 경우로, 식 (2.1), (2.2), (2.3) 모형이 다음과 같다.

$$
\begin{aligned}
\hat{\theta}_{it} &= \theta_{it} + e_{it}, \quad t = 1, \ldots, T; \;\; i = 1, \ldots, m, \\
\theta_{it} &= \boldsymbol{z}_{it}^T\boldsymbol{\beta} + v_i + u_{it} \\
&= \rho\theta_{i,t-1} + (\boldsymbol{z}_{it} - \rho\boldsymbol{z}_{i,t-1})^T\boldsymbol{\beta} + (1-\rho)v_{i+u_{it}}, \quad \text{(AR(1) 모형인 경우)}, \\
\hat{\theta}_{it} &= \boldsymbol{z}_{it}^T\boldsymbol{\beta} + v_i + u_{it} + e_{it}.
\end{aligned}
$$

미국, 캐나다의 소지역 고용통계 모형이 TSCS 모형에 기반을 두고 연구되고 있다.

2.2. 경험적 베이즈(EB) 추정량

식 (2.3)의 Fay-Harriot 모형을 베이지안 형식으로 요약하면 $\hat{\theta}_i|\theta_i \sim i.i.d.\ N(\theta_i, V_i)$, $\theta_i \sim i.i.d.\ N(\boldsymbol{z}_i^T\hat{\boldsymbol{\beta}}, \sigma_v^2)$이다. 이 모형에서 θ_i의 사후분포는

$$\theta_i|\hat{\theta}_i \sim i.i.d.\ N\left[\alpha_i\hat{\theta}_i + (1-\alpha_i)\boldsymbol{z}_i^T\hat{\boldsymbol{\beta}},\ \alpha_i V_i\right]$$

이다. 이 모형에서 평균제곱오차(MSE)를 최소화하는 경험적 베이즈 추정량(Empirical Bayes Estimator; EB 추정량)을 얻을 수 있다. EB 추정량은 다음과 같다.

$$\hat{\theta}_i^{EB} = \hat{\alpha}_i\hat{\theta}_i + (1-\hat{\alpha}_i)\boldsymbol{z}_i^T\hat{\boldsymbol{\beta}}.$$

$\hat{\theta}_i$의 주변분포는 $\hat{\theta}_i \sim iid\ N(\boldsymbol{z}_i^T\hat{\boldsymbol{\beta}},\ \sigma_v^2 + V_i)$이고, $\hat{\alpha}_i$, $\hat{\beta}$, $\hat{\sigma}_v^2$은 다음과 같다 (Fay와 Harriot, 1979).

$$
\begin{aligned}
\hat{\alpha}_i &= \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \hat{V}_i} \\
\hat{\beta} &= \left(\sum_{i=1}^m \frac{\boldsymbol{z}_i\boldsymbol{z}_i^T}{\hat{\sigma}_v^2 + \hat{V}_i}\right)^{-1}\left(\sum_{i=1}^m \frac{\boldsymbol{z}_i\hat{\theta}_i}{\hat{\sigma}_v^2 + \hat{V}_i}\right) \\
\hat{\sigma}_v^2 &= \left(\sum_i \left(\hat{\theta}_i - \boldsymbol{z}_i^T\boldsymbol{\beta}^*\right)^2 - \sum_i \hat{V}_i h_{ii}\right)\Big/(m-p)
\end{aligned}
$$

여기서, $\hat{V}_i$는 직접추정량 $\hat{\theta}_i$에 대한 분산 추정량, m은 소지역의 수, p는 보조변수의 수, $h_{ii} = \boldsymbol{z}_i^T(\sum_i \boldsymbol{z}_i\boldsymbol{z}_i^T)^{-1}\boldsymbol{z}_i$, $\beta^* = (\sum_i \boldsymbol{z}_i\boldsymbol{z}_i^T)^{-1}(\sum_i \boldsymbol{z}_i\hat{\theta}_i)$이다.

Prasad와 Rao (1990)는 EB 추정량에 대한 평균제곱오차를 근사적으로 계산하는 방법을 제시했다.

$$
\begin{aligned}
\text{mse}\left(\hat{\theta}_i^{EB}\right) &= g_{1i}\left(\hat{\sigma}_v^2\right) + g_{2i}\left(\hat{\sigma}_v^2\right) + 2g_{3i}\left(\hat{\sigma}_v^2\right) \\
g_{1i}\left(\hat{\sigma}_v^2\right) &= \alpha_i\hat{V}_i \\
g_{2i}\left(\hat{\sigma}_v^2\right) &= \hat{\sigma}_v^2(1-\alpha_i)^2 z_i^T\left(\sum_i \alpha_i z_i z_i^T\right)^{-1} z_i
\end{aligned}
$$

$$g_{3i}\left(\hat{\sigma}_v^2\right) = \hat{V}_i^2\left(\hat{\sigma}_v^2 + \hat{V}_i\right)^{-3} h\left(\hat{\sigma}_v^2\right), \quad h\left(\hat{\sigma}_v^2\right) = 2m^{-2}\sum_i\left(\hat{\sigma}_v^2 + \hat{V}_i\right)^2.$$

Jiang Lahiri와 Wan (2002)는 EB 추정량의 MSE 추정에 대한 잭나이프 방법을 제안하였다. 이 방법은 범주형 자료에 대한 소지역모형에도 쉽게 적용이 가능한 일반적인 방법으로 알려져 있다. MSE 추정을 위한 잭나이프 절차는 다음과 같다.

1단계: 전체 표본자료에 대한 $\hat{\sigma}_v^2$, $\hat{\beta}$, $\hat{\theta}_i^{EB}$와 k번째 지역의 자료 $(\hat{\theta}_k, z_k)$을 제거한 후의 $\hat{\sigma}_v^2(k)$, $\hat{\beta}(k)$, $\hat{\theta}_i^{EB}(k)$, $(k = 1, \ldots, m)$를 계산.

2단계: $g_{1i}(\hat{\sigma}_v^2)$의 편향 조정된(bias-adjusted) 잭나이프 추정치 $\hat{M}_{1i}$를 계산

$$\hat{M}_{1i} = g_{1i}\left(\hat{\sigma}_v^2\right) - \frac{m-1}{m}\sum_{k=1}^{m}\left(g_{1i}\left(\hat{\sigma}_v^2(k)\right) - g_{1i}\left(\hat{\sigma}_v^2\right)\right).$$

3단계: $\hat{M}_{2i}$를 계산

$$\hat{M}_{2i} = \frac{m-1}{m}\sum_{k=1}^{m}\left(\hat{\theta}_i^{EB}(k) - \hat{\theta}_i^{EB}\right)^2.$$

4단계: 잭나이프 MSPE 추정량을 계산

$$\mathrm{mse}_J\left(\hat{\theta}_i^{EB}\right) = \hat{M}_{1i} + \hat{M}_{2i}.$$

$\hat{M}_{1i}$는 모형의 모수를 안다고 가정할 때 MSE의 추정량이며, $\hat{M}_{2i}$는 모형 모수의 추정으로부터 발생하는 추가적인 변동을 추정한다. 잭나이프 MSE 추정량은 근사적 불편추정량이다.

3. 시군별 실업자수 추정

경제활동인구조사는 1,629개 조사구내에서 20가구를 추출하여 15세 이상 가구원을 대상으로 조사를 실시하고 있다. 전체 표본규모는 전국 약 32,000여 가구원이며, 매월 전국과 16개 시도에 대한 경제활동인구수, 취업자수, 실업자수, 비경제활동인구수를 작성하고 있다. 전국 기준 취업자수, 경제활동인구수, 비경제활동인구수의 상대표준오차는 0.5%내외이나 실업자수의 상대표준오차는 3%내외로 매우 높은 수준이다. 또한 16개 시도별 실업자수의 월별 상대표준오차는 5~25%로 현재의 표본규모로는 시군별 실업자수를 작성할 수 없는 상황이다.

이에 본 논문에서는 소지역추정기법을 적용하여 시군 실업자수를 추정하고자 한다. 본 논문에서 사용되는 소지역모형은 Fay-Harriot의 기본모형이며, EB 추정방법에 의해 실업자수를 추정하고자 한다. EB 추정은 미지의 초모수를 최우추정법, 적률법 혹은 UMVUE 등을 사용하여 추정하며 추정된 사전분포를 사용하는 방법으로 계층적 베이즈(HB) 추정방법과 함께 소지역추정을 위해 널리 사용되고 있다. 추정을 위해 2005년 11월의 경제활동인구조사 자료를 이용하였으며, 소지역모형을 위한 보조자료로 센서스 및 인구특성분포와 관련된 행정자료를 적용하였다.

EB 추정의 가정인 모형의 정규성 가정을 만족하기 위하여 실업자수에 대한 로지스틱 변환을 실시하였으며 평균제곱오차(MSE)에 의해 각 결과들을 비교하였다. 경제활동인구조사의 시군별 표본은 표본조사구수가 4개 이하를 포함하고 있는 시군이 전체 158개 시군(특광역시 제외)의 65%를 차지하고 있다. 이 중 표본조사구수가 5개 이상(100가구 이상)인 55개 시군만을 분석에 포함하였다.

표 1: 성별, 연령대별 인구비율, 농가비율과 실업자수간 상관분석

		주민등록인구비율(15세 이상)					농업총조사
		남자	15~30세	30~45세	45~65세	65세이상	농가비율
실업자수(인구	r	0.16135	0.38706	0.59740	−0.48458	−0.61007	−0.62465
총조사 10%표본)	p-value	0.2221	0.0025	<.0001	0.0001	<.0001	<.0001

3.1. 모형기반추정을 위한 보조변수 선택

소지역모형을 위한 보조정보는 가능하면 표본오차가 없는 센서스나 행정자료를 이용해야 한다. 이용 가능한 행정자료로는 주민등록자료, 고용보험 및 실업수급 등이 있으나, 고용보험 및 실업수급에 대한 행정자료는 고용보험 의무가입 사업체 및 실업수급자 대상이 크게 확대되어 안정화되기 전에는 고용관련 정부정책 변화에 따라 모형이 영향을 받을 가능성이 존재한다. 향후 실업수당에 대한 정부정책이 안정화 된 이후 실업관련 DB를 시군단위로 연계한다면 시군별 고용통계 중 실업자수를 위한 좋은 보조정보로 사용하는 것이 효율적일 것이다. 이에 본 논문에서 보조변수로는 주민등록인구자료 및 센서스를 보조정보로 활용하고자 한다.

1차로 고려된 보조변수는 2005년 인구총조사 10% 표본조사에서 파악된 시군별 15세 이상 인구의 취업자 비율과 실업자 비율이다. 2차로는 2005년 기준 주민등록상 남녀별 인구비율과 연령대별(15~30세, 30~45세, 45~65세, 65세 이상) 인구비율을 보조변수로 고려하였다. 3차로는 농가비율을 고려하였는데 2005년 농업총조사의 결과 농가비율이다. 2005년 이후의 농가비율에 대해서는 조사시점 기준의 농업조사 결과를 활용할 수 있다. 농업조사의 경우 추정오차가 존재하지만 농업조사의 표본규모가 6만가구 이상으로 시군별 농가수에 대한 상대표준오차가 대부분 시군에 대해서 5%내외로 본 모형에서는 농가비율에 대한 추정오차를 무시할 수 있을 것이다.

성별 및 연령대별 인구특성과 농가비율이 실업자수와 어느 정도의 상관관계를 갖는지를 파악하기 위해 Pearson 상관분석을 실시하였다. 상관분석 결과 실업자수는 성별과는 유의미한 관계를 보이지 않으며, 15~30세, 30~45세 인구비율과는 양의 상관관계, 45~65세 및 65세 이상 인구비율과 농가비율과는 음의 상관관계를 갖는다.

3.2. 시군 실업자수 추정을 위한 소지역모형

모형을 위한 주요 관심모수는 전체 실업자수 Y_i로 정하고, 보조변수는 주민등록 상 연령그룹별(15~30세, 30~45세, 45~65세, 65세 이상) 인구수 z_1, z_2, z_3, z_4, 인구주택총조사에서 15세 이상 인구의 취업자 비율과 실업자 비율 z_5, z_6, 농업총조사에서 농가비율 z_7로 정한다.

지역수준 랜덤효과 모형과 EB 추정량은 다음과 같다.

$$\hat{\theta}_i = \boldsymbol{z}_i^T\boldsymbol{\beta} + v_i + e_i, \quad i = 1, 2, \ldots, m, \;\; v_i \sim N\left(0, \sigma_v^2\right), \;\; e_i \sim N\left(0, \hat{V}_i\right),$$
$$\hat{\theta}_i^{EB} = \hat{\alpha}_i\hat{\theta}_i + (1 - \hat{\alpha}_i)\boldsymbol{z}_i^T\hat{\boldsymbol{\beta}}.$$

EB 추정량을 위해 관심변수인 실업자수는 정규분포를 따라야 한다. 그러나, $\theta_i = Y_i$일 때 그림 1로 부터 θ_i는 정규성 가정을 만족하지 못하므로 변수변환을 고려해야 한다.

3가지 모형을 고려하자. 모형1은 실업자수, 모형2는 실업자수를 15세 이상 인구수로 나눈 실업비율, 모형3은 실업자수에 대한 로짓변환한 값을 각각 관심모수를 한 Fay-Harriot 모형이다. 세가지 모수를 다음과 같이 정의하자.

$$\theta_{1i} = Y_i, \qquad \theta_{2i} = \frac{Y_i}{N_i}, \qquad \theta_{3i} = \log\left(\frac{\theta_{2i}}{1 - \theta_{2i}}\right).$$

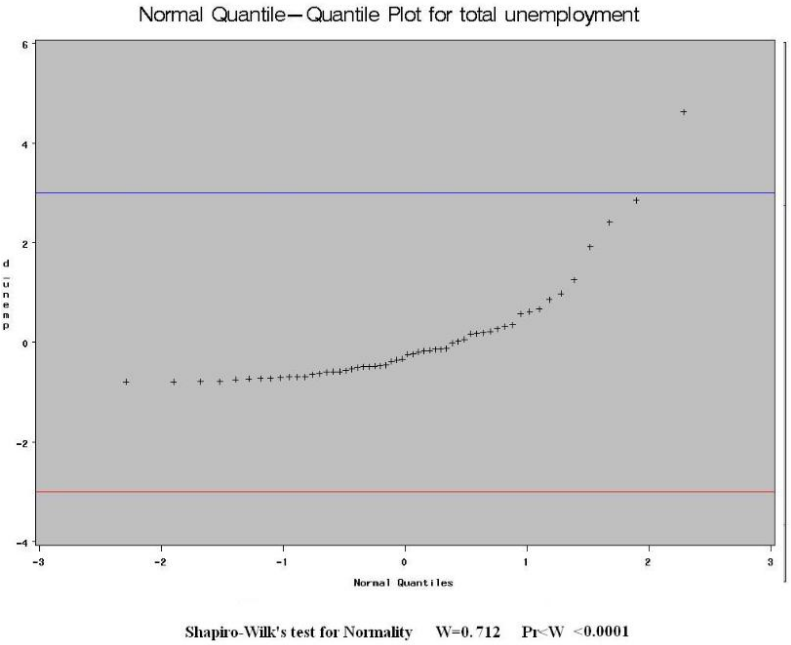


그림 1: *실업자수($\hat{\theta}_{1i}$)의 Q-Q 플롯*

따라서, 각 모수에 대한 직접추정량과 EB 추정량은 다음과 같다.

$$\hat{\theta}_{1i} = \hat{Y}_i, \qquad \hat{\theta}_{1i}^{EB} = \hat{\alpha}_{1i}\hat{\theta}_{1i} + (1-\hat{\alpha}_{1i})\boldsymbol{z}_i^T\hat{\boldsymbol{\beta}}_1, \tag{3.1}$$

$$\hat{\theta}_{2i} = \frac{\hat{Y}_i}{N_i}, \qquad \hat{\theta}_{2i}^{EB} = \hat{\alpha}_{2i}\hat{\theta}_{2i} + (1-\hat{\alpha}_{2i})\boldsymbol{z}_i^T\hat{\boldsymbol{\beta}}_2, \tag{3.2}$$

$$\hat{\theta}_{3i} = \log\left(\frac{\hat{\theta}_{2i}}{1-\hat{\theta}_{2i}}\right), \qquad \hat{\theta}_{3i}^{EB} = \hat{\alpha}_{3i}\hat{\theta}_{3i} + (1-\hat{\alpha}_{3i})\boldsymbol{z}_i^T\hat{\boldsymbol{\beta}}_3. \tag{3.3}$$

그림 1~3으로부터 $\hat{\theta}_{1i}$은 정규성 가정을 만족하지 못하며, $\hat{\theta}_{2i}$와 $\hat{\theta}_{3i}$은 정규성 가정을 만족함을 알 수 있다.

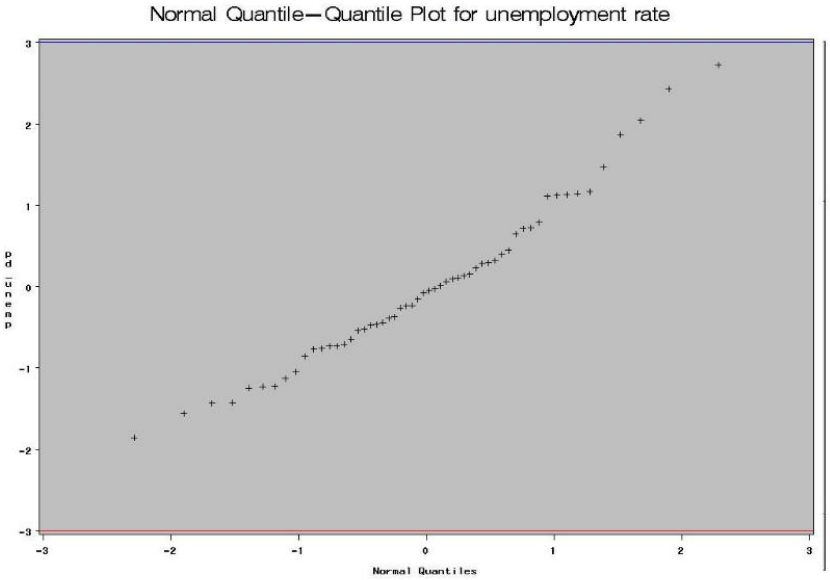
3.3. 추정결과

식 (3.1)~(3.3)에 의해 추정된 EB 추정결과는 표 2와 같다. $\hat{Y}_i$, $\widehat{\mathrm{Var}}(\hat{Y}_i)$, $\widehat{\mathrm{RSE}}(\hat{Y}_i)$는 직접추정량, 직접추정량에 대한 분산 및 상대표준오차 추정량이다.

$$\widehat{\mathrm{RSE}}\left(\hat{Y}_i\right) = \frac{\sqrt{\widehat{\mathrm{Var}}\left(\hat{Y}_i\right)}}{\hat{Y}_i} \times 100.$$

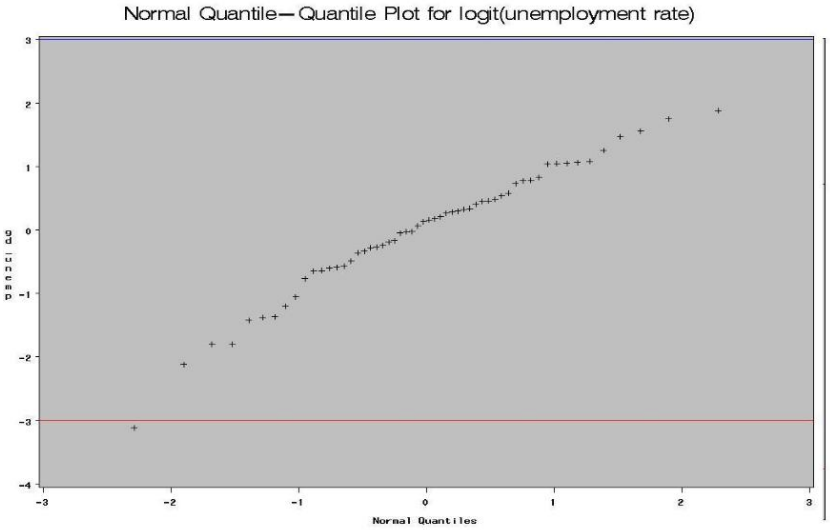
$\hat{\theta}_i^{EB}$, $\widehat{\mathrm{mse}}(\hat{\theta}_i^{EB})$, $\widehat{\mathrm{RMSE}}(\hat{\theta}_i^{EB})$는 EB 추정량과 EB 추정량에 대한 분산 및 상대표준오차 추정량이다. $\widehat{\mathrm{mse}}(\hat{\theta}_i^{EB})$는 Jiang 등 (2002)의 잭나이프 방법에 의해 구하였다.

$$\widehat{\mathrm{RMSE}}\left(\hat{\theta}_i^{EB}\right) = \frac{\sqrt{\widehat{\mathrm{mse}}\left(\hat{\theta}_i^{EB}\right)}}{\hat{\theta}_i^{EB}} \times 100.$$



Shapiro-Wilk's test for Normality W=0.969, Pr<W =0.1557

그림 2: *실업비율($\hat{\theta}_{2i}$)의 Q-Q 플롯*



Shapiro-Wilk's test for Normality W=0.973, Pr<W =0.2226

그림 3: *실업자수의 로짓변환($\hat{\theta}_{3i}$)의 Q-Q 플롯*

표 2는 2005년 11월 경제활동인구조사 결과자료를 활용하여 55개 시군의 실업자수를 추정한 결과이다. 모형3은 모형에 의한 오차가 적으므로 RMSE가 가장 적게 산출되었다. 따라서 모형3은 세 모형 중에서 시군별 실업자수의 추정량에 대한 신뢰도와 정확도측면에서 좋은 모형임을 알 수 있다.

표 2: EB 추정량 결과비교

시군 일련번호	표본 조사구수	직접 추정량		모형1 (θ_{1i}에 대한 모형)		모형2 (θ_{2i}에 대한 모형)		모형3 (θ_{3i}에 대한 모형)	
		$\hat{\theta}_{1i}$	RSE	$\hat{\theta}_{1i}^{EB}$	RMSE	$\hat{\theta}_{2i}^{EB}$	RMSE	$\hat{\theta}_{3i}^{EB}$	RMSE
1	16	28,924	23.2%	17,590	14.0%	23,715	19.2%	27,598	21.0%
2	15	19,638	32.6%	14,923	19.9%	16,677	25.4%	17,697	26.6%
3	6	5,905	57.3%	5,285	42.3%	5,710	31.5%	5,763	35.3%
4	10	6,575	36.5%	7,676	23.9%	7,856	25.0%	7,805	28.5%
5	13	14,760	22.7%	13,168	18.1%	14,358	19.5%	14,573	20.5%
6	6	6,411	41.4%	4,669	44.1%	5,065	28.4%	5,485	30.5%
7	8	9,262	34.4%	6,442	32.9%	5,331	29.9%	6,723	27.5%
8	9	17,328	31.9%	11,176	23.0%	13,635	22.6%	15,861	26.4%
9	13	11,322	42.0%	13,106	19.8%	13,315	26.6%	13,138	32.6%
10	11	3,917	56.6%	4,856	35.9%	4,838	34.7%	5,044	36.4%
11	5	5,658	73.5%	5,609	42.4%	6,515	28.3%	6,622	39.2%
12	10	8,228	29.8%	6,787	28.4%	8,630	23.6%	8,675	24.9%
13	5	2,234	39.2%	2,125	40.0%	2,137	31.4%	2,172	30.4%
14	5	1,171	50.0%	1,195	48.3%	1,347	36.4%	1,505	34.1%
15	6	1,807	63.3%	2,300	46.8%	2,506	36.9%	2,857	36.6%
16	5	2,199	44.8%	2,035	46.3%	1,857	34.1%	1,969	31.6%
17	13	2,161	28.3%	2,196	27.4%	2,288	24.6%	2,408	24.5%
18	15	2,262	41.5%	2,426	37.1%	2,574	30.7%	2,756	30.6%
19	13	608	70.7%	663	64.5%	896	45.3%	1,743	33.9%
20	5	864	61.3%	821	63.8%	1,023	36.4%	1,071	36.4%
21	5	965	66.7%	933	67.7%	689	49.5%	777	39.7%
22	30	5,713	26.5%	6,826	19.8%	6,682	20.8%	6,800	23.3%
23	12	2,283	31.6%	2,278	30.9%	2,175	27.8%	2,260	25.8%
24	8	1,479	42.3%	1,439	42.7%	1,498	32.3%	1,561	30.8%
25	26	7,976	19.7%	7,832	18.1%	7,501	18.3%	7,777	18.3%
26	7	1,679	40.4%	1,597	41.6%	1,334	36.0%	1,469	30.2%
27	7	1,163	45.6%	1,137	46.0%	1,052	36.7%	1,106	32.0%
28	9	3,760	35.7%	3,361	36.8%	2,337	34.9%	2,801	28.6%
29	8	2,405	43.2%	2,265	43.5%	1,494	40.2%	1,719	32.2%
30	6	1,978	74.0%	1,701	77.0%	1,234	49.0%	1,425	37.3%
31	5	1,687	45.5%	1,671	44.9%	1,081	45.6%	1,232	34.0%
32	22	5,766	33.7%	7,060	23.3%	6,373	26.8%	6,534	28.1%
33	10	3,847	72.8%	3,181	61.0%	2,648	44.8%	2,990	38.2%
34	11	4,039	34.2%	3,915	32.0%	3,529	30.1%	3,760	27.8%
35	5	953	70.7%	941	70.2%	862	49.2%	965	37.6%
36	14	4,095	23.8%	3,789	24.6%	3,154	23.5%	3,612	21.2%
37	16	3,520	27.3%	3,467	26.6%	3,208	25.2%	3,326	23.4%
38	13	2,757	49.2%	2,803	44.1%	2,677	33.8%	2,777	32.7%
39	5	603	70.7%	607	69.8%	668	49.3%	803	38.3%
40	6	1,633	70.7%	1,518	71.3%	1,331	42.0%	1,412	39.5%
41	17	5,078	31.3%	5,662	24.8%	4,955	28.2%	5,060	26.2%
42	11	1,419	50.9%	1,605	44.1%	1,659	39.6%	2,094	33.9%
43	6	989	70.7%	1,003	68.1%	1,102	46.8%	1,254	37.8%
44	8	2,934	41.8%	2,605	43.8%	1,911	37.7%	2,221	30.4%
45	13	4,018	32.2%	4,286	27.8%	4,656	23.5%	4,594	26.9%
46	5	630	100.0%	645	95.9%	906	49.8%	1,137	40.0%
47	6	1,076	46.8%	1,080	46.2%	906	44.2%	998	33.8%
48	8	3,013	44.9%	2,862	43.2%	2,866	31.8%	3,032	32.6%

시군 일련번호	표본 조사구수	직접 추정량		모형1 (θ_{1i}에 대한 모형)		모형2 (θ_{2i}에 대한 모형)		모형3 (θ_{3i}에 대한 모형)	
		$\hat{\theta}_{1i}$	RSE	$\hat{\theta}_{1i}^{EB}$	RMSE	$\hat{\theta}_{2i}^{EB}$	RMSE	$\hat{\theta}_{3i}^{EB}$	RMSE
49	16	9,879	30.8%	7,968	26.2%	8,470	23.5%	9,147	26.2%
50	13	6,219	34.6%	5,755	30.0%	5,228	30.3%	5,501	27.8%
51	10	3,496	68.7%	3,636	50.0%	3,447	40.3%	3,655	37.1%
52	10	7,757	25.1%	6,908	24.3%	6,988	21.2%	7,360	22.3%
53	6	4,879	70.2%	2,910	77.5%	2,989	34.0%	3,265	37.3%
54	29	4,726	24.7%	4,812	22.7%	4,562	22.7%	4,604	22.1%
55	16	675	43.5%	683	42.9%	761	36.9%	980	33.0%

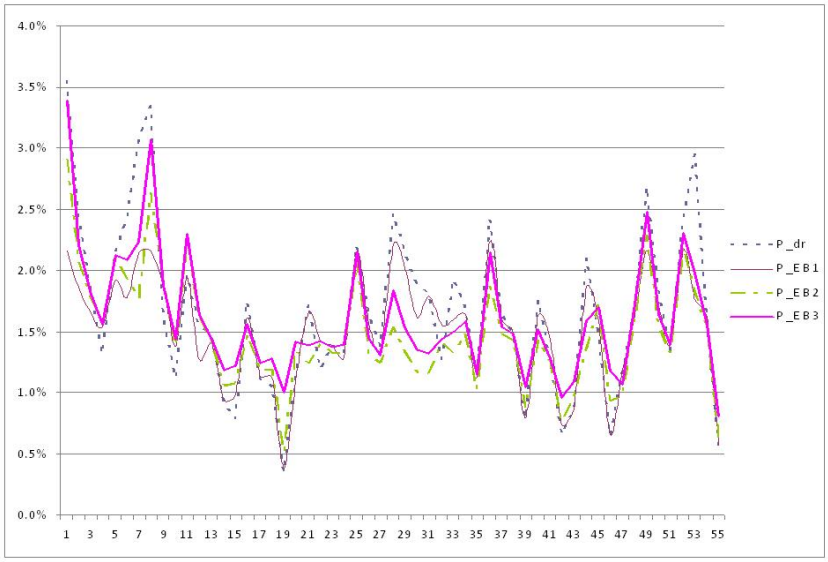


그림 4: *각 모형별 실업비율 비교*

그림 4는 각 모형별로 추정된 실업자수를 15세 이상 인구수로 나눈 실업비율을 나타낸다. 표 2에서의 결과에서도 알 수 있듯이, $\hat{\theta}_{3i}^{EB}$에 의한 실업비율이 직접추정량과 전반적으로 가깝고, $\hat{\theta}_{2i}^{EB}$에 의한 실업비율이 다음으로 가까움을 알 수 있다. 몇몇 시군에서의 $\hat{\theta}_{1i}^{EB}$가 직접추정량과 가까운 경향을 보이기도 하나, 해당 시군의 표본규모가 작으며 분산추정량 또한 커서 신뢰도가 낮음을 알 수 있다.

4. 결론

본 논문에서는 시군단위의 실업자수를 추정하고자 하였다. 이를 위해 센서스 실업자 및 취업자수, 연령대별 인구 그리고 농가수를 보조변수로 채택하였고, 관심변수인 실업자수에 대해 로짓변환을 실시하였다. 또한 Fay-Harriot 기본모형을 바탕으로 EB 추정량에 의해 55개 시군의 실업자수를 추정하였다. 그러나 100가구 이하의 표본규모를 갖으며 표본가구가 전혀 없는 시군도 포함하고 있는 나머지 103개 시군에 대한 소지역 모형에 대한 연구가 더 필요하다.

통계조사 실무자와 최종 이용자에게 모형의 선택, 특히 보조변수의 선택이 매우 중요하다. 모형에 대한 가정을 잘 검토하고, 실업자수와 같이 정규분포 등 가정을 만족하지 않는 경우 적적한 변수변환을 통하여 추정량의 정도를 향상시켜야 한다. 또한 행정자료와 센서스 자료는 표본오차를 포함하지 않기

때문에 이런 자료를 보조변수로 사용하는 것이 추정량의 오차측면에서 유리하다. 그러나, 행정자료와 전수자료에서 발생할 수 있는 비표본오차를 주의해야 한다.

경제활동인구조사 자료와 보조변수간 조사단위별로 연결이 가능한 단위수준모형(Unit level model)이 지역수준모형(Area level model)보다 더 효율적일 수 있다. 그러한 경우 지분오차에 의한 선형회귀모형(Nested error linear regression model)이 가능하다. 그러나 연결이 현실적으로 거의 불가능할 것이다.

최근 복합 소지역모형에 대한 MSE(특히, $\hat{M}_{2i}$) 추정에 대한 연구가 활발하게 진행되고 있으며 시군별 실업자수에 대해서도 MSE에 대한 연구가 지속적으로 진행되어야 한다.

참고 문헌

이계오 (2000). 시군구 실업자 추정을 위한 소지역 추정법, <응용통계연구>, **13**, 275–286.

이상은 (2001). Bayes prediction for small area estimation, <한국통계학회논문집>, **8**, 407–416.

이상은 (2006). 공간통계량을 활용한 베이지안 자기 포아송 모형을 이용한 소지역 통계, <응용통계연구>, **19**, 421–430.

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

통계청 (2002). <소지역통계 추정법>, Unpubliched Report.

통계청 (2004). <모형기반 소지역추정방법>, Unpubliched Report.

Battese, G. F., Harter, R. M. and Fuller, W. A. (1988). An error component model for prediction of county crop area using survey and satellite data, *Journal of American Statistical Association*, **83**, 28–36.

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.

Fay, R. E. and Harriot, R. A. (1979). Estimation of income of small places: An application of James-Stein procedures to census data, *Journal of the American Statistical Association*, **74**, 269–271.

Gonzalez-Manteiga, W., Lombardia, M. J., Molina, I., Morales, D. and Santa-Maria, L. (2006). Bootstrap mean squared error of a small-area EBLUP, *Journal of Statistical Computation and Simulation*, **78**, 443–462.

Hall, P. and Maiti, T. (2006). Nonparametric estimation of mean squared prediction error in nested-error regression models, *Annals of Statistics*, **34**, 1733–1750.

Jiang, J., Lahiri, P. and Wan, S. M. (2002). A unified Jackknife theory for empirical best prediction with M-estimation, *Annals of Statistics*, **30**, 1782–1810.

Opsomer, J. D., Claeskens, G., Ranalli, M. G., Kauermann, G. and Breidt, F. J. (2008). Non-parametric small area estimation using penalized spline regression, *Journal of the Royal Statistical Society B*, **70**, 265–286.

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

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

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

Wang, J. and Fuller, W. A. (2003). The mean squared error of small area predictors constructed with estimated error variances, *Journal of the American Statistical Association*, **98**, 716–723.

Ybarra, L. M. R. and Lohr, S. L. (2008). Small area estimation when auxiliary information is measured with error, *Biometrika*, **95**, 919–931.

2010년 1월 접수; 2010년 4월 채택

Small Area Estimation to Unemployment Statistics in Korea

Jin Kim[1,a], Jae-kwang Kim[b]

[a]Sampling Division, Statistics Korea; [b]Deparatment of Statistics, Iowa State University

Abstract

Most sample surveys are designed to estimate reliable statistics for the whole population and for some large subpopulations. However, the research for small area estimation have been increasing in recent years because users demand to reliable estimates for smaller subpopulations like small areas or specific domains. In Korea, the Economically Active Population Survey(EAPS) is the main household survey that produces monthly unemployment rates for nationwide and 16 large areas (7 metropolitans and 9 provinces) in Korea. For county level estimation, direct estimators are not reliable because of the small sample sizes. We consider small area estimation of the county level unemployment ratesfrom the sample observations in EAPS. To do this, we use an area level model to "borrow strength" from the auxiliary information, such as administrative data and census data. The proposed method is based on the assumption of normality of the model errors in the area level model. The proposed method is compared with the other alternatives in terms of the estimated mean squared errors.

Keywords: Area level model, borrow strength, indirect estimation, model-based estimation.

[1] Corresponding author: Assistant Director, Sampling Division, Statistics Korea, Building Ⅲ,Government Complex-Daejeon 139,Seonsaro, Seo-gu, Daejeon 302-701, Korea. E-mail: jinknso@korea.kr