소지역 추정법을 이용한 시군구의 실업자 추정

이 계 오,* 정 연 수**

<요 약>

신뢰할 만한 소지역 통계 작성을 위한 다양한 소지역 추정 기법들이 최근 많은 관심속에 개발되고 있다. 이 논문은 다양한 소지역 추정 기법들 중 일부 기법들에 대한 간략한 소개 및 실례를 제시한다. 먼저 대표적인 소지역에 대한 간접추정법인 인구통계학적 방법, 합성추정법과 복합추정법에 관한 이론 및 추정절차를 살펴보았고, 모형 기반 추정법으로써 경험적 베이즈(EB) 추정법과 계층적 베이즈(HB) 추정법을 소개하였다. 마지막으로 합성추정법과 복합추정법을 이용하여 충북의 시군구 실업자 추정에 적용해 보았고, 시군구 실업자 추정결과를 직접 추정법의 결과와 비교하였다.

Ⅰ. 서 론

지리적으로 구분된 영역 또는 좀 더 일반적으로 임의의 부분모집단이 소지역으로 간주될 수 있다. 지리적으로 구분된 도 단위가 소지역이 될 수도 있고, 도 단위 내의 시군구 단위가 소지역이 될 수도 있다. 통상적으로 소지역에 대한 표본의 크기는 소지역의 크기에 따라 증가하는 경향이 있으나, 대영역을 기준으로 작성된 표본설계에서 소지역을 추정할 경우에는 반드시 이러한 경향을 보이는 것은 아니다. 소지역에 대한 추정값들은 정부의 지역별 예산 배분등의 정책 입안 시에 절대적인 참고자료로 활용될 수 있기 때문에 최근에는 신뢰할 수 있는 소지역 통계 작성에 많은 관심들을 보이고 있다.

5년 또는 10년 마다 실시되는 센서스 자료, 실업자의 구직 등록 자료나 실직 보험 자료와

* (363-849) 충북 청원군 남일면 쌍수리 사서함 335-2호 공군사관학교 전산통계학과 교수
e-mail) kayolee@hanimail.com
** (363-849) 충북 청원군 남일면 쌍수리 사서함 335-2호 공군사관학교 전산통계학과 조교수
e-mail) ys_chung@hanimail.com

같은 행정보고 자료, 다양한 사회적 관심에 의해 실시되는 표본조사 자료 등이 소지역 추정을 위해 활용된다. 그러나 많은 경우 표본조사에 의해 확보된 직접 추정값들은 대영역에 대해서는 신뢰할만한 수준이나, 소지역에서는 확보된 표본이 작을 경우가 대부분이므로 신뢰할만한 정확도를 기대하기는 어려운 실정이다. 이러한 이유 때문에 유사한 특성을 갖는 인근 지역으로부터 정보를 취득하여 소지역의 추정의 정확도를 높일 수 있는 간접 추정방법을 생각하게 되는데, 이러한 추정방법이 소지역 추정에 활용된다.

간접 추정량은 최근의 센서스 자료라든가 행정보고 자료와 같은 보조자료를 소지역들과 연계할 수 있는 암시모형(Implicit Model) 또는 명시모형(Explicit Model)을 설정하여 추정된다. 암시모형 하에서 추정되는 간접 추정량으로 합성 추정량(Synthetic Estimator)과 복합 추정량(Composite Estimator)을 들 수 있고, 모형에 근거한 추정량으로는 경험적 베이즈(EB) 추정량, 계층적 베이즈(HB) 추정량 등을 들 수 있다. 암시모형을 이용한 소지역 추론은 오랫동안 인구통계학자들에 의해 진행되어 왔고, 최근에 들어서는 모형 기반 추정법들에 관한 연구 논문들이 발표되고 있다. 소지역 추정방법에 관한 제반 기법들과 적용상의 문제점들은 Ghosh and Rao(1994), Rao(1999)을 참고하면 자세히 소개되어 있다.

소지역 통계 작성의 요구에 부응하여 여러 심포지움과 워크숍들이 개최되어, 다양한 소지역 추정기법과 응용관련 논문들이 발표되고 있다(National Institute on Drug Abuse (Princeton,1979), International Simposium on Small Area Statistics(Ottawa, 1987), International Scientific Conference on Small Area Statistics and Survey Designs Warsawa, 1992), International Association of Survey Statisticians Satellite Conference on Small Area Estimation (Riga,1999)).

이 논문에서는 소지역 추정기법에 관한 일부 이론과 적용예를 간략히 소개하기로 한다. 먼저 대표적인 간접 추정법인 인구통계학적 방법, 합성 추정법과 복합 추정법을 살펴보고, 모형기반 추정법으로는 경험적 베이즈(EB) 추정법과 계층적 베이즈(HB) 추정법을 간략히 소개하며, 충북 지역에 대해 위의 합성추정법과 복합추정법을 적용하여 시군구 실업자 추정에 적용해 보기로 한다.

Ⅱ. 직접 추정법

경활조사에서 해당 시군구에 배정되어 조사된 조사구들 만을 이용하여 해당 시군구의 실업자 수를 추정하는 형식이다. i 시군구의 실업자 수에 대한 추정량은 조사 모집단과 표본간의 관계에서 산출한 승수 ξ와 관찰된 자료 y들의 일차결합으로 다음과 같이 표현할 수 있다.

$$\hat{Y}_{i\cdot} = \sum_{j \in s_i} \xi_j y_j. \qquad (2.1)$$

여기에서 s_i 는 i 시군구에서 조사한 조사구들의 집합이다.

식 (2.1)에서 주어진 추정량은 불편추정량이 되도록 ξ_j 를 산정하나 해당 시도내의 시군구별 조사구 수의 불균형적 분포로 추정량의 변동도 불균형적 분포를 갖기 때문에 추정량의 신뢰성 조정에 문제가 있을 수 있다.

Ⅲ. 간접 추정법(Indirect Estimation)

3.1 인구통계학적 방법(Demographic Method)

미국의 경우에서와 같이 10년 주기로 센서스를 할 경우, 지방 도시나 county의 중간 해당 년도의 인구를 추정하기 위해서 사용하는 추정법으로 센서스 자료와 인구수에 관련된 징후 변수(출생자수, 사망자수, 주택 수, 등록한 학생 수 등)의 변동을 분석하여 얻은 예측값을 결합하는 추정법을 인구 통계학적 방법이라 말한다.

3.1.1 생멸률법(Vital Rates Method: VR Method)

VR법은 출생과 사망에 관련된 자료를 이용하여 인구의 변동률보다는 징후 변수의 영향만을 분석하여 활용한다. 가장 최근에 센서스를 실시한 해를 기준 년도로 하고, 기준해로부터 t 년 후에 소지역의 인구수를 추정하고자 한다. 여기에서 전제 조건은 추정 대상인 소지역을 포함하는 대지역의 특성과 소지역의 특성이 동일하다는 것이며, 전제 조건에서 많이 벗어나는 경우에는 추정량의 편향이 커져서 신뢰도가 낮아진다.

t 년 후의 소지역의 출생률과 사망률을 γ_{bt} 와 γ_{dt} 로 표현하고 대영역의 출생률과 사망률을 R_{bt} 와 R_{dt} 라 나타내면 다음과 같은 관계가 주어진다.

$$\gamma_{bt} = \gamma_{bo}\left(\frac{R_{bt}}{R_{bo}}\right), \qquad \gamma_{dt} = \gamma_{do}\left(\frac{R_{dt}}{R_{do}}\right) \qquad (3.1)$$

여기에서 γ_{b0} 와 γ_{d0} 는 기준 해의 소지역의 출생률과 사망률이고 R_{b0}와 R_{d0}는 기준 해의 대지역의 출생률과 사망률을 의미한다.

센서스를 실시한 기준해로부터 t 년 후의 인구 수는 다음 식에 의해서 추정할 수 있다.

$$\underline{p}_t = \frac{1}{2}\left(\frac{b_t}{\gamma_{bt}} + \frac{d_t}{\gamma_{dt}}\right) \qquad (3.2)$$

단, b_t와 d_t는 소지역의 t년 후의 출생자수와 사망자수를 뜻한다.

3.1.2 성분법(Components Method)

성분법은 출생과 사망 인구수 및 유입, 유출 인구에 관한 자료를 이용하여 소지역의 인구수를 추정하기 위해 고안된 방법이다. 센서스를 실시한 기준해로부터 t 년 동안의 출생 인구, 사망인구 및 총 이주인구를 각각 $b_{0,t}$, $d_{0,t}$, $m_{0,t}$로 나타냈을 때 t년 후의 인구수는 다음식에 의해 추정한다.

$$p_t = p_0 + b_{0,t} - d_{0,t} + m_{0,t} \qquad (3.3)$$

여기에서 $m_{0,t} = i_{0,t} - e_{0,t} + n_{0,t}$ 로 계산하며, $i_{0,t}$ 는 유입인구, $e_{0,t}$ 는 유출인구, $n_{0,t}$ 는 주 간의 총 이주인구를 나타내며 행정보고자료에 의해 주어진다.

3.1.3 회귀 징후법(Regression Symptomatic Procedures)

회귀 징후법은 다중선형회귀모형을 이용하여 소지역의 인구를 추정하는 방법으로써 징후변수들을 독립변수로 선택하여 소지역 추정에 이용한다. 비 상관계수(Ratio Correlation), 차분상관계수(Difference Correlation), 표본 회귀법(Sample Regression Method) 등은 이러한 회귀징후법의 일종이다. 여기에서는 다른 두 방법보다는 비교적 자주 사용되고 있는 표본 회귀법을 설명하기로 한다. 먼저 종속변수와 독립변수를 다음과 같이 정의하자.

$Y_i = (p_{it}/P_t)/(p_{i0}/P_0) =$ i 소지역의 인구비 변화량,

$x_{ij} = (s_{ijt}/S_{jt})/(s_{ij0}/S_{j0}) =$ i 소지역에 대한 j 번째 징후변수 s_j의 변화량,

여기에서 P_t , P_0 , S_{jt} , S_{j0}는 i 소지역을 포함하는 대지역에서의 값들이고, x_{ij} 는 행정자료로부터 얻는다($j = 1, 2, \cdots, p$).

회귀 표본법은 종속변수 Y_i 가 징후변수 $x_{i1}, x_{i2}, \cdots, x_{ip}$ 의 일차결합으로 표현될 수 있다는 것을 가정하며, 이때 Y_i 의 값은 조사된 직접추정값 $\hat{Y}_i$ 을 이용하여 m 개의 소지역 중 k개의 소지역에 대하여 선형회귀식을 적합시켜 회귀계수들을 추정한 후, Y_i 의 추정값으로 다음의 표본회귀 추정량을 이용한다.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip} , \quad i = 1, 2, \cdots, m \qquad (3.4)$$

i 소지역에 대한 인구수는 (3.4)식의 표본 회귀추정량을 이용하여 다음식으로 추정한다.

$$\tilde{p}_{it} = \hat{Y}_i (p_{i0}/P_0) \hat{P}_t , \quad i = 1, 2, \cdots, m \qquad (3.5)$$

표본 회귀추정량은 표본으로부터 직접 추정된 값이 아니라 다중 선형회귀를 거쳐 얻어진 보정된 추정량이며, 표본 회귀법은 이를 이용하여 소지역의 인구를 추정하는 방법이다. 그러

나 이러한 방법은 추후 논의될 모형에 근거한 소지역 추정보다는 효율성이 상당히 떨어지는 것으로 밝혀지고 있다.

3.2 합성 추정법(Synthetic Estimation)

추정하고자 하는 소지역과 특성이 유사한 소지역들의 정보를 이용하여 추정값의 정도를 높이고자하는 추정방식을 합성 추정법이라하며, 주변이나 유사지역의 정보를 이용하므로 "Borrow Strength"라고 말하기도 한다. 표본조사의 설계 시에는 대영역에 대해서만 직접 추정값을 구하고자 하였으나 대영역을 분할한 소지역의 추정값이 필요한 때에는 대영역과 소지역의 구조적 특성이 같다는 조건하에서 소지역의 연구변수에 대한 추정값을 구할 수 있는데, 이때 대영역의 분할은 지리적인 분할보다는 연령대별 또는 교육정도별과 같은 특성에 따른 분할을 말한다.

대영역을 I개 소지역으로 분할하며 또한 대영역을 특성 기준에 따라 J개의 범주로 분류한다면 i소지역의 추정값은 다음 식으로 구할 수 있다.

$$\hat{Y}_{i\cdot} = \sum_{j}^{J} \hat{p}_{ij}\ \hat{Y}_{\cdot j} \qquad (3.6)$$

단, $\hat{p}_{ij}$는 i번째 소지역의 j범주에 대한 가중값이며 센서스나 행정자료에서 구해진다. $Y_{\cdot j}$는 대영역에서 j범주에 대한 표본 추정값이다. 단, 대영역의 표본의 수는 충분하게 많아서 신뢰성 있는 추정값을 구할 수 있다고 가정한다. i소지역의 실업자 추정에 관한 경우를 생각해 보자.

Y_{ij} : i소지역의 j범주(연령대별 또는 교육정도별)의 실업자수,

X_{ij} : i소지역의 j범주(연령대별 또는 교육정도별)의 경제활동인구,

$\hat{Y}_{\cdot j} = \sum_{i} Y_{ij}$: j범주의 대영역에 대한 합계,

$Y_{i\cdot} = \sum_{j} Y_{ij}$: i소지역의 실업자 수.

$\hat{Y}_{\cdot j}$의 직접 추정값 $\hat{Y}_{d\cdot j}$는 표본조사 자료만으로 추정가능하고, X_{ij}는 센서스 또는 행정자료 등 보조변수의 정보에서 계산 가능한 것으로 가정한다면 합성추정량은 다음과 같이 나타낼 수 있다.

$$\hat{Y}_{i\cdot}^{\ s} = \sum_{j} \left(\frac{X_{ij}}{X_{\cdot j}} \right)\ \hat{Y}_{d\cdot j} \qquad (3.7)$$

만약 $\hat{Y}_{d\cdot j}$가 비 추정량의 형식을 갖는다면, $\hat{Y}_{d\cdot j} = (\hat{Y}_{\cdot j} / \hat{X}_{\cdot j}) X_{\cdot j}$로 나타낼 수 있으므로 (3.7)식은 다음과 같이 표현될 수 있다.

$$\hat{Y}_{i.}^{s} = \sum_{j} X_{ij}\left(\frac{\hat{Y}_{.j}}{\hat{X}_{.j}}\right) = \sum_{j}\left(\frac{X_{ij}}{\hat{X}_{.j}}\right)\hat{Y}_{.j} \qquad (3.8)$$

여기에서 $\hat{Y}_{i.}^{s}$ 가 불편추정량이 되기 위해서는 $\frac{Y_{.j}}{X_{.j}} = \frac{Y_{ij}}{X_{ij}}$ 를 만족해야 하고, 이를 만족하지 못할 경우에는 편향추정량이 되며, 이때 $\hat{Y}_{i.}^{s}$의 편향의 크기는 $B(\hat{Y}_{i.}^{s}) = E(\hat{Y}_{i.}^{s} - Y_{i.})$이다. 즉, $B(\hat{Y}_{i.}^{s}) = \sum_{j} X_{ij}(\frac{Y_{.j}}{X_{.j}} - \frac{Y_{ij}}{X_{ij}})$.

$\hat{Y}_{i.}^{s}$의 평균제곱오차($MSE(\hat{Y}_{i.}^{s})$)의 근사적 불편추정량은 다음과 같이 주어질 수 있다.

$$\widehat{MSE}(\hat{Y}_{i.}^{s}) = (\hat{Y}_{i.}^{s} - \hat{Y}_{i.})^2 - \widehat{Var}(\hat{Y}_{i.}) \qquad (3.9)$$

3.3 복합 추정법(Composite Estimation)

소지역에 배정된 표본수가 적기 때문에 표본 조사만을 이용한 직접 추정량의 불안정에서 오는 낮은 신뢰성과 합성추정량의 편향을 보완하기 위해서 직접 추정값과 합성 추정값의 가중평균은 사용하는데 이를 복합 추정량(Composite Estimator)이라 한다.

$$\hat{Y}_{i.}^{c} = w_i\,\hat{Y}_{i.} + (1-w_i)\,\hat{Y}_{i.}^{s} \qquad (3.10)$$

여기에서 $\hat{Y}_{i.}$는 표본조사에서 직접 계산한 추정값이며, $\hat{Y}_{i.}^{s}$는 합성 추정값을 나타낸다. w_i는 가중값으로 0 과 1 사이의 값이다.

먼저 평균제곱오차 $MSE(\hat{Y}_{i.}^{c})$를 최소화하는 w_i는 아래와 같다.

$$w_{i(opt)} = \frac{MSE(\hat{Y}_{i.}^{s})}{MSE(\hat{Y}_{i.}^{s}) + V(\hat{Y}_{i.})} \qquad (3.11)$$

최적 가중값 $w_{i(opt)}$의 추정값은 다음 식으로 계산된다.

$$\hat{w}_{(opt)} = \frac{mse(\hat{Y}_{i.}^{s})}{(\hat{Y}_{i.}^{s} - \hat{Y}_{i.})^2} \qquad (3.12)$$

모든 소지역에 공통 가중값을 부여하는 방법으로써 초기 공통 가중값 w를 이용하여 $MSE(\hat{Y}_{i.}^{s})$들의 평균을 최소화하는 가중값은 아래와 같다.

$$\hat{w}_{(opt)} = 1 - \frac{\sum_{i} \hat{V}(\hat{Y}_{i.})}{\sum_{i}(\hat{Y}_{i.}^{s} - \hat{Y}_{i.})^2} \qquad (3.13)$$

각 소지역에 배정된 표본 크기에 의존하는 가중값은 다음과 같이 계산된다.

$$w_i(\delta) = \begin{cases} 1, & \hat{N}_i \geq \delta N_i \\ \dfrac{\hat{N}_i}{\delta N_i}, & (\text{그외}) \end{cases} \tag{3.14}$$

단, N_i는 i소지역의 크기이며 $\hat{N}_i = N(n_i/n)$이다. $\hat{N}_i$는 직접추정량이며 δ는 합성추정량의 기여도를 조정하는 값이므로 주관적으로 결정한 값이다. 예를 들어 캐나다 노동력 통계조사에서는 2/3으로 한다.

어떤 추정법에 의해서 소지역의 추정값을 구하더라도 대영역을 소지역으로 분할하여 각 소지역의 추정값을 추정하므로 소지역의 추정값의 합계는 대영역의 추정값과 같아야 할 것이다. 왜냐하면 매월 정부기관에서 발표하는 광역시와 도의 실업자수와 해당 소지역의 추정값의 합계가 같도록 조정하지 않으면 서로 상이한 통계수치로 인하여 혼란을 줄 수 있기 때문에 한 가지 통계수치가 되도록 조정된 추정량을 계산해야 할 것이다. 각 소지역의 추정량을 생명률법, 합성추정법 또는 복합 추정법 등의 어느 한 방법으로 계산한 것으로 간주할 때 조정된 소지역 추정량은 다음과 같다.

$$\hat{Y}_{i.}^{A} = \left(\frac{\hat{Y}_{i.}^{*}}{\sum_i \hat{Y}_{i.}^{*}}\right) \hat{Y} \tag{3.15}$$

단, $\hat{Y}$는 광역시·도의 직접 추정값이며, $\hat{Y}_{i.}^{*}$는 i소지역을 *추정법으로 추정한 것이다.

Ⅳ. 모형 기반 추정법

4.1 기본적인 지역 수준 모형

소지역 추정시 모형에 근거한 추정방법이 많은 사람들의 관심을 끌고 있는 것은 다음과 같은 몇가지 장점에 기인한다. 먼저 모형 기반 추정법은 소지역들을 연결하고 있는 모형 구조가 소지역 간의 복잡한 오차구조를 내포하고 있기 때문에 소지역 간의 변동을 반영하여 소지역 추정의 정확도를 높일 수 있다는 점이며, 또한 표본자료로부터 모형의 유용성이 확인될 수 있고, 연속형의 자료뿐만 아니라 범주형 자료 및 시계열 자료와 같은 다양한 경우들에 대해서도 모형화하여 추론할 수 있으며, 모형 기반 추정법으로 소지역 추정량들과 연관있는 많은 측도들이 얻어질 수 있다는 장점들을 들 수 있다.

지역 간의 공변량을 포함하고 있는 지역 수준 모형을 이용하여 경험적 최량선형불편예측(EBLUP) 추정량, 경험적 베이즈(EB) 추정량, 계층적 베이즈(HB) 추정량에 대해 설명하기로

한다. 지역 수준 모형은 기본적으로 두가지의 성분들로 이루어진다. 즉, 소지역에 대한 직접 추정량 $\hat{\theta}_i$ 과 소지역의 보조변수들로 표현되는 θ_i 의 두 가지 성분들을 모형으로 연결하여 모형 기반 추정량을 찾아내게 된다.

지정된 함수 $g(\cdot)$에 대하여 직접 추정량 $\hat{\theta}_i = g(\hat{\bar{Y}}_i)$은 모집단의 값 $\theta_i = g(\bar{Y}_i)$와 표본추출오차 e_i 에 의해 다음과 같이 표현될 수 있다.

$$\hat{\theta}_i = \theta_i + e_i \quad , \quad i = 1, 2, \cdots, m \tag{4.1}$$

여기에서 표본추출오차 e_i 는 서로 독립이며, 평균이 0, 분산이 ψ_i 임이 가정되며, 보통 ψ_i 는 기지로 가정된다.

θ_i 는 소지역의 정보를 나타내는 보조변수 $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \cdots, z_{ip})^T$를 이용하여 선형회귀모형을 통해 표현한다.

$$\begin{aligned} \theta_i &= z_{i1}\beta_i + z_{i2}\beta_2 + \cdots + z_{ip}\beta_p + v_i \\ &= \boldsymbol{z}_i^T \boldsymbol{\beta} + v_i \end{aligned} \tag{4.2}$$

여기에서 모형오차 v_i 는 서로 독립이며, 평균이 0, 분산 σ_v^2을 갖고, 표본추출오차 e_i 와는 서로 독립임을 가정한다.

마지막으로 (4.1)과 (4.2)의 두 성분들을 결합하면 다음과 같은 결합모형을 얻을 수 있다.

$$\hat{\theta}_i = \boldsymbol{z}_i^T \boldsymbol{\beta} + v_i + e_i \tag{4.3}$$

위의 결합모형은 고정효과 $\boldsymbol{\beta}$와 소지역 랜덤효과 v_i를 갖는 선형혼합효과모형의 일종이며, 특히 설계 기반 확률변수(design-based random variable) e_i와 모형 기반 확률변수(model-based random variable) v_i를 동시에 포함하고 있는 모형이다. 여기에서 모수 σ_v^2은 소지역들의 동질성을 나타내는 측도이다.

4.2 경험적 베이즈(EB)방법

경험적 최량선형불편예측(EBLUP) 방법, 경험적 베이즈(EB) 방법 및 계층적 베이즈(HB) 방법은 모형에 근거한 소지역 추정문제에 많이 활용되고 있는 방법이다. 특히 경험적 최량선형불편예측 방법은 선형혼합모형을 이용한 추론에 응용되어 왔고, 경험적 베이즈 방법 및 계층적 베이즈 방법은 좀 더 일반적인 모형을 이용한 소지역 추정에 활용되고 있다.

EBLUP 추정량은 랜덤오차 e_i와 v_i의 분포에 대한 가정을 필요로 하지 않으나, *MSE* 추정을 위해 정규분포를 가정하기도 한다. 또한, EBLUP 추정량과 EB 추정량은 e_i와 v_i

를 정규분포로 가정했을 경우에는 동일하며, HB 추정량과는 근사적으로 같게 나타난다. 그러나 추정량들의 변동을 나타내는 측도들은 동일하지는 않다.

고정계수 l_t를 갖는 θ_i 의 선형추정량 $\sum l_t \hat{\theta}_t$ 가 모형 (4.3)에 대해서 $\sum l_t \hat{\theta}_t - \theta_i$ 의 기대값이 0 을 만족할 때, $\sum l_t \hat{\theta}_t$ 를 θ_i 의 선형불편예측(LUP) 추정량이라 한다. θ_i 의 최량선형불편예측(BLUP) 추정량은 선형불편예측(LUP) 추정량들 중 최소평균제곱오차를 갖는 추정량을 말한다.

모형 (4.3)하에서 θ_i 의 BLUP 추정량은 다음과 같이 주어진다(Prasad and Rao,1990).

$$\tilde{\theta}_i(\sigma_v^2) = \gamma_i \hat{\theta}_i + (1-\gamma_i)\ \boldsymbol{z}_i^T\ \tilde{\boldsymbol{\beta}}(\sigma_v^2) \qquad (4.4)$$

여기에서 $\gamma_i = \sigma_v^2/(\sigma_v^2 + \psi_i)$이고, $\tilde{\boldsymbol{\beta}}(\sigma_v^2)$은 가중치 $(\sigma_v^2 + \psi_i)^{-1}$을 갖는 가중최소제곱추정량으로 아래와 같이 주어진다.

$$\tilde{\boldsymbol{\beta}}(\sigma_v^2) = (\sum_i \gamma_i\ \boldsymbol{z}_i\ \boldsymbol{z}_i^T)^{-1}(\sum_i \gamma_i\ \boldsymbol{z}_i y_i) \qquad (4.5)$$

(4.4)식의 BLUP 추정량은 가중치 γ_i를 갖는 직접추정량 $\hat{\theta}_i$ 과 가중치 $1-\gamma_i$를 갖는 회귀합성추정량 $\boldsymbol{z}_i^T \tilde{\boldsymbol{\beta}}(\sigma_v^2)$의 가중결합으로 볼 수 있다. 또한, 표본분산 ψ_i가 작을때 (σ_v^2이 클 경우) BLUP 추정량은 직접추정량 $\hat{\theta}_i$ 에 큰 가중치가 부여되고, 반대의 경우에는 회귀합성추정량 $\boldsymbol{z}_i^T \tilde{\boldsymbol{\beta}}(\sigma_v^2)$에 큰 가중치가 부여된다. 표본이 추출되지 않은 지역들에 대해서는 BLUP 추정량은 회귀합성추정량만으로 주어질 수 있다.

BLUP 추정량의 변동의 측도는 추정량의 MSE ($= E(est. - \theta_i)^2$)에 의해 주어지며 다음과 같다.

$$MSE\{\tilde{\theta}_i(\sigma_v^2)\} = g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2) \qquad (4.6)$$

여기에서 $g_{1i}(\sigma_v^2) = \gamma_i \psi_i$ 이고,

$g_{2i}(\sigma_v^2) = \sigma_v^2(1-\gamma_i)^2\ \boldsymbol{z}_i^T(\sum_i \gamma_i\ \boldsymbol{z}_i\ \boldsymbol{z}_i^T)^{-1}\ \boldsymbol{z}_i$ 로 주어진다. 식(4.4)와 (4.6)은 랜덤오차 v_i와 e_i에 관한 분포의 가정을 필요로 하지는 않는다.

주요 항 $g_{1i}(\sigma_v^2) = \gamma_i \psi_i$는 $O(1)$, $g_{2i}(\sigma_v^2)$은 $O(m^{-1})$에 유계인 항이며, 이로부터 BLUP 추정량의 MSE 값은 γ_i 나 모형분산 σ_v^2이 표본분산 ψ_i 에 비해 작을 경우 직접추정량의 MSE 값보다 훨씬 작아질 수 있다는 사실을 알 수 있다. 따라서 소지역 추정의 정확도는 표본분산에 비해 모형분산을 작게할 수 있는 보조변수에 크게 의존한다고 볼 수 있다.

대부분의 문제에서는 모형분산 σ_v^2은 미지이므로 적절한 $\widehat{\sigma_v^2}$을 추정하여 EBLUP 추정량 $\widetilde{\theta_i} = \widetilde{\theta_i}(\widehat{\sigma_v^2})$을 산출한다. 이때 소지역의 평균 $\overline{Y_i}$ 의 추정량은 $g^{-1}(\widetilde{\theta_i})$ 로, σ_v^2의 추정량은 $\widehat{\sigma_v^2} = \max(\widetilde{\sigma_v^2}, 0)$로 주어진다. 여기에서 $\widetilde{\sigma_v^2}$은 다음 식을 만족한다.

$$(m-p)\ \widetilde{\sigma_v^2} = \sum_i (\widehat{\theta_i} - \boldsymbol{z}_i^T \boldsymbol{\beta}^*)^2 - \sum_i \psi_i h_{ii} \qquad (4.7)$$

(4.7)식에서 $h_{ii} = \boldsymbol{z}_i^T (\sum_i \boldsymbol{z}_i \boldsymbol{z}_i^T)^{-1} \boldsymbol{z}_i$ 이고, $\boldsymbol{\beta}^*$는 $\boldsymbol{\beta}$의 *OLS*(ordinary least squares) 추정량이다. 한편, $\widetilde{\sigma_v^2}$은 다음과 같은 비선형 방정식의 반복적인 해로써 구할 수도 있다.

$$a(\sigma_v^2) = \sum_i \left\{ \widehat{\theta_i} - \boldsymbol{z}_i^T \widetilde{\boldsymbol{\beta}}(\sigma_v^2) \right\}^2 / (\sigma_v^2 + \psi_i) = m-p \qquad (4.8)$$

여기에서 $\widetilde{\boldsymbol{\beta}}(\sigma_v^2)$은 (4.5)식에 주어졌고, (4.8)식의 가운데 항은 가중잔차제곱합, $m-p$는 가중잔차제곱합과 관계가 되는 자유도이다. 만약 $\widehat{\sigma_v^2} = 0$ 이면 EBLUP 추정량 $\widetilde{\theta_i}$ 는 회귀합성추정량 $\boldsymbol{z}_i^T \widehat{\boldsymbol{\beta}}$ 로 축약된다. 단, $\widehat{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}(\widehat{\sigma_v^2})$이며, 식(4.5)에서 σ_v^2 대신에 $\widehat{\sigma_v^2}$을 대체하여 산출한다. 물론 위의 (4.7), (4.8)식으로부터 얻게되는 추정량들도 v_i와 e_i의 분포에 대한 가정을 필요로 하지는 않는다.

만약 랜덤오차 v_i와 e_i가 정규분포를 따른다고 가정한다면, $\widehat{\theta_i}$는 평균이 $\boldsymbol{z}_i^T \boldsymbol{\beta}$ 이고 분산이 $\sigma_v^2 + \psi_i$ 인 서로 독립인 정규분포를 따르게 된다. 이러한 분포에 대한 가정하에서 계산된 $\boldsymbol{\beta}$ 와 σ_v^2의 최대우도추정량을 제한최대우도추정량(REML)이라 하며, 선형혼합모형하에서도 근사적으로 유효하다. 따라서 $\widetilde{\theta_i}$ 의 BLUP 추정량을 산출 시 σ_v^2의 REML 추정량을 이용하여도 근사적으로 타당하다.

경험적 베이즈(EB) 추정법은 랜덤오차 v_i와 e_i가 정규분포를 따른다는 가정하에서 출발한다. $(\widehat{\theta_i}, \theta_i)$의 결합분포가 평균이 $(\boldsymbol{z}_i^T \boldsymbol{\beta}, \boldsymbol{z}_i^T \boldsymbol{\beta})$, 분산이 $(\sigma_v^2 + \psi_i, \sigma_v^2)$, 상관계수가 γ_i 인 이변량 정규분포를 따른다고 가정하자. 이때, θ_i 의 평균제곱오차를 최소로 하는 베이즈 추정량은 다음과 같다.

$$\widetilde{\theta_i}^B(\boldsymbol{\beta}, \sigma_v^2) = E(\theta_i | \widehat{\theta_i}, \boldsymbol{\beta}, \sigma_v^2) = \gamma_i \widehat{\theta_i} + (1-\gamma_i)\ \boldsymbol{z}_i^T \boldsymbol{\beta} \qquad (4.9)$$

(4.9)식의 베이즈 추정량은 선형성 또는 불편성을 만족하지는 않는다. 여기에서 모수 $\boldsymbol{\beta}$와 σ_v^2을 제한최대우도(REML) 추정량으로 대체하여 다음과 같은 θ_i 의 경험적 베이즈

(EB) 추정량을 얻는다.

$$\tilde{\theta}_i^{EB} = \tilde{\theta}_i^{B}(\hat{\beta}, \hat{\sigma}_v^2) \tag{4.10}$$

경험적 베이즈(EB) 추정량 $\tilde{\theta}_i^{EB}$ 는 정규분포의 가정하에서는 EBLUP 추정량 $\tilde{\theta}_i$ 와 같다. 그러나 경험적 베이즈방법은 $\tilde{\theta}_i$ 과 θ_i 의 임의의 결합분포에 대해서도 일반적으로 응용할 수 있다는 점을 장점으로 들 수 있다.

EBLUP 추정량 $\tilde{\theta}_i = \tilde{\theta}_i(\hat{\sigma}_v^2)$의 MSE 추정량은 (4.6)식에서 σ_v^2 대신 $\hat{\sigma}_v^2$을 대체하여 얻어질 수 있으나, 이 경우에는 σ_v^2에 대한 추정효과가 무시되기 때문에 MSE의 추정값은 과소추정되는 경향을 보인다. 이러한 문제 때문에 Prasad and Rao(1990)는 $\{v_i\}$와 $\{e_i\}$에 대해 정규성을 가정하여 근사적으로 불편인 EBLUP 추정량 $\tilde{\theta}_i$ 의 MSE 추정량을 제안하였다. Prasad and Rao(1990)가 제안한 MSE 추정량은 (4.7)식의 σ_v^2의 적률추정량을 사용하였을 경우 다음과 같이 주어진다.

$$mse(\tilde{\theta}_i) = g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_v^2) \tag{4.11}$$

여기에서 $g_{1i}(\sigma_v^2) = \gamma_i\psi_i$, $g_{2i}(\sigma_v^2) = \sigma_v^2(1-\gamma_i)^2 \boldsymbol{z}_i^T(\sum_i \gamma_i \boldsymbol{z}_i \boldsymbol{z}_i^T)^{-1}\boldsymbol{z}_i$, $g_{3i}(\sigma_v^2) = \{\psi_i^2/(\sigma_v^2+\psi_i)^3\}h(\sigma_v^2)$, $h(\sigma_v^2) = 2m^{-2}\sum_i(\sigma_v^2+\psi_i)^2$로 주어진다.

최근들어 Jing, Lahiri and Wan(1999)은 근사적으로 불편인 젝나이프 MSE 추정량을 제안하였다. 젝나이프 방법은 랜덤인 지역효과들을 갖는 로지스틱 회귀와 같은 좀 더 복잡한 모형들에 대해서도 쉽게 적용할 수 있다는 장점을 갖고 있다.

θ_i 의 EB 추정량 (4.10)을 $\tilde{\theta}_i^{EB} = k(\hat{\theta}_i, \hat{\varphi})$로 표현할 때, 젝나이프 절차는 다음과 같다. 여기에서 $\varphi = (\beta, \sigma_v^2)$은 모형에서의 모수 β 와 σ_v^2 을 나타낸다.

(i) l 번째 지역의 자료 $(\hat{\theta}_l, \boldsymbol{z}_l)$을 제외한 φ 의 추정량 $\hat{\varphi}(l)$을 계산한다.

이때의 EB 추정량을 $\tilde{\theta}_i^{EB}(l) = k(\hat{\theta}_i, \hat{\varphi}(l))$로 나타내자.

(ii) $\hat{M}_{2i} = \frac{m-1}{m}\sum_{l=1}^{m}(\tilde{\theta}_i^{EB}(l) - \tilde{\theta}_i^{EB})$ 를 계산한다.

(iii) $\hat{M}_{1i} = g_{ii}(\hat{\sigma}_v^2) - \frac{m-1}{m}\sum_{l=1}^{m}\{g_{ii}(\sigma_v^2(l) - g_{1i}(\hat{\sigma}_v^2)\}^2$ 을 계산한다.

(iv) 마지막으로 MSE 의 젝나이프 추정량 $mse_J(\tilde{\theta}_i^{EB}) = \hat{M}_{1i} + \hat{M}_{2i}$를 계산한다.

$\hat{M}_{1i}$ 은 φ 가 기지일 때 MSE에 대한 추정량이며, $\hat{M}_{2i}$ 는 모형 모수 φ 를 추정할 때 추가적으로 발생하는 MSE 에 대한 변화량을 추정한다.

4.3 계층적 베이즈(HB) 방법

계층적 베이즈(HB) 방법을 이용한 추론은 비교적 추론의 정확도가 높고, 복잡한 유형의 문제들에서도 최근에 개발된 MCMC(Monte Carlo Markov Chain)방법을 이용하여 해결할 수 있다. 깁스 샘플러가 이러한 방법의 일종이다. HB 방법에서는 모형 모수 $\varphi = (\beta, \sigma_v^2)$뿐만 아니라 모집단의 값 θ_i 가 랜덤으로 간주되며, 모형 모수들에 대한 사전분포가 명시된다. θ_i 의 추론은 주변 사후분포에 의해 결정된다.

즉, 주어진 자료 $\{(\hat{\theta}_i, z_i), i=1, 2, \cdots, m\}$에 대한 조건부 분포 $f(\theta_i | \hat{\theta})$에 의해 추론이 행해진다. 여기에서 $\hat{\theta}$ 은 직접추정값 $\hat{\theta}_i$의 벡터이다. 특히 θ_i 는 사후분포의 평균 $E(\theta_i | \hat{\theta})$에 의해 추정되며, 추정량의 변동은 사후분포의 분산 $V(\theta_i | \hat{\theta})$에 의해 추정된다.

먼저 σ_v^2 이 기지인 상태를 가정하고 β 에 관한 사전분포를 배정하기로 한다. β의 사전분포가 상수에 비례하고(i.e improper prior), v_i 와 e_i 가 정규분포를 따른다고 가정한다면, 이때 사후평균 $E(\theta_i | \hat{\theta}, \sigma_v^2)$은 (4.4)식의 BLUP 추정량 $\tilde{\theta}_i(\sigma_v^2)$과 동일하다. 더욱이 사후분산 $V(\theta_i | \hat{\theta}, \sigma_v^2)$은 (4.6)식의 BLUP 추정량의 MSE 와 같다. 따라서 σ_v^2이 기지인 상태에서는 HB 방법과 EBLUP 방법은 동일한 추론을 이끌어 낸다고 볼수 있다.

실제의 문제에서는 σ_v^2은 대부분 미지의 값으로 나타난다. 이러한 경우에는 β 뿐만 아니라 σ_v^2에 관한 사전분포를 고려해야 하며, 또한 서로 독립임을 가정하여 주변사후분포 $f(\sigma_v^2 | \hat{\theta})$을 이끌어 낸다. 만약 σ_v^2에 관한 사전분포를 불완전(improper) 사전분포를 배정한다면, θ_i 의 사후분포가 불완전 사후분포가 될 수 있기 때문에 이러한 문제를 피하기 위해서 $\tau_v = \sigma_v^{-2}$의 사전분포를 gamma(a, b)와 같이 배정한다 (여기에서 gamma(a, b) : $f(\tau_v) \propto \exp(-a\tau_v)\ \tau_v^{b-1}$). 주변사후분포 $f(\sigma_v^2 | \hat{\theta})$를 이용한 HB 추정량 $E(\theta_i | \hat{\theta})$은 다음 식과 같이 주어진다.

$$\tilde{\theta}_i^{HB} = E(\theta_i \mid \hat{\boldsymbol{\theta}}) = \int \tilde{\theta}_i(\sigma_v^2) f(\sigma_v^2 \mid \hat{\boldsymbol{\theta}})\, d\sigma_v^2 \qquad (4.12)$$

위의 (4.12)식을 $E_{\sigma_v^2 \mid \hat{\boldsymbol{\theta}}}\{\tilde{\theta}_i(\sigma_v^2)\}$으로 표현하면, 사후분산 $V(\theta_i \mid \hat{\boldsymbol{\theta}})$ 은 다음과 같다.

$$V(\theta_i \mid \hat{\boldsymbol{\theta}}) = E_{\sigma_v^2 \mid \hat{\boldsymbol{\theta}}}\{g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2)\} + V_{\sigma_v^2 \mid \hat{\boldsymbol{\theta}}}\{\tilde{\theta}_i(\sigma_v^2)\} \qquad (4.13)$$

여기에서 $V_{\sigma_v^2 \mid \hat{\boldsymbol{\theta}}}$ 은 $f(\sigma_v^2 \mid \hat{\boldsymbol{\theta}})$에 관한 분산을 의미한다.

위에서 소개한 (4.12)와 (4.13)은 일차원 수치적분으로 계산된다. 좀 더 복잡한 모형에 대한 고차원 수치적분은 MCMC 방법을 이용하여 계산할 수 있다. (4.12)식으로부터 $\tilde{\theta}_i^{HB}$는 EBLUP(EB) 추정량 $\tilde{\theta}_i(\hat{\sigma}_v^2)$와 근사적으로 같다는 것을 알 수 있다.

깁스 샘플링은 위의 (4.12)와 (4.13)을 결정하기 위해 사용될 수 있는 일종의 MCMC 방법이다. 깁스 샘플링을 수행하기 위해서는 다음과 같은 깁스 조건부 분포들이 필요하다.

(i) $\boldsymbol{\beta} \mid \boldsymbol{\theta}, \sigma_v^2, \hat{\boldsymbol{\theta}} \sim N_p((\sum \boldsymbol{z}_i \boldsymbol{z}_i^T)^{-1}(\sum \boldsymbol{z}_i \theta_i), \sigma_v^2(\sum \boldsymbol{z}_i \boldsymbol{z}_i^T)^{-1})$

(ii) $\theta_i \mid \boldsymbol{\beta}, \sigma_v^2, \hat{\boldsymbol{\theta}} \sim N(\tilde{\theta}_i^B(\boldsymbol{\beta}, \sigma_v^2), g_{1i}(\sigma_v^2) = \gamma_i \psi_i)$

(iii) $\tau_v = \sigma_v^{-2} \mid \boldsymbol{\beta}, \boldsymbol{\theta}, \hat{\boldsymbol{\theta}} \sim \text{gamma}(\tilde{a}, \tilde{b})$,

단, $\tilde{a} = \frac{1}{2}\sum(\theta_i - \boldsymbol{z}_i^T)^2 + a$, $\tilde{b} = \frac{m}{2} + b$.

깁스 앨고리즘은 다음 절차에 의해 이루어진다.

(a) $\theta_i = \theta_i^{(0)}$, $\sigma_v^2 = \sigma_v^{2(0)}$을 초기값으로 하여 (i)로부터 $\boldsymbol{\beta}^{(1)}$을 계산

(b) $\boldsymbol{\beta} = \boldsymbol{\beta}^{(1)}$, $\sigma_v^2 = \sigma_v^{2(0)}$를 이용하여 (ii)로부터 $\theta_i^{(1)}$, $i = 1, 2, \cdots, m$ 을 계산

(c) $\theta_i = \theta_i^{(1)}$과 $\boldsymbol{\beta} = \boldsymbol{\beta}^{(1)}$을 이용하여 (iii)으로부터 $\sigma_v^{2(1)}$을 계산

(d) 절차 (a), (b), (c)를 한 사이클로 하여 반복 수행

수렴이 이루어지는 시점 t 까지 충분히 반복한 후, 이 후부터 얻어지는 J 개의 표본 $\{\boldsymbol{\beta}^{(t+j)}, \sigma_v^{2(t+j)}, \theta_1^{(t+j)}, \cdots, \theta_m^{(t+j)}; j = 1, 2, \cdots, J\}$을 $\boldsymbol{\beta}, \sigma_v^2, \theta_1, \cdots, \theta_m$의 결합 사후분포로 얻은 표본으로 간주한다. 초기값은 보통 $\theta_i^{(0)} = \tilde{\theta}_i^{EB}$, $\sigma_v^{2(0)} =$

σ_v^2의 REML 추정량을 사용한다.

위에서 계산된 J 개의 표본을 이용하여 θ_i 의 사후평균, 사후분산을 다음과 같이 추정한다.

$$\hat{\theta}_i^{HB} \approx \frac{1}{J}\sum_j \hat{\theta}_i(\sigma_v^{2(t+j)}) = \frac{1}{J}\sum_j \hat{\theta}_i(j) = \hat{\theta}_i(\cdot) \quad (4.14)$$

$$V(\theta_i|\hat{\theta}) \approx \frac{1}{J}\sum_j \{g_{1i}(\sigma_v^{2(t+j)}) + g_{2i}(\sigma_v^{2(t+j)})\} + \frac{1}{J}\sum_j \{\hat{\theta}_i(j) - \hat{\theta}_i(\cdot)\}^2 \quad (4.15)$$

V. 시·군·구의 실업통계 개발 예

5.1 개 요

충북의 행정구역은 2구 2시 8군으로 편성되어 있다. 다음의 <표1>은 1999년 4월의 경제활동인구 조사 결과를 요약한 것이다.

<표1> 충북의 경제활동인구 총괄

(단위 : 천명, %)

구 분	15세이상인구	경제활동인구	비경제활동인구	경제활동참가율
남	501	361	140	72.06
여	570	287	283	50.35
전체	1,071	648	423	60.50

충북의 경제활동 참가율이 남자의 경우, 여자에 비해 비교적 높은 것으로 나타나고 있다. 다음의 <표2>는 충북지역의 1999년 4월 시와 군 지역을 구분하여 성별에 따른 경제활동 인구와 조사구 수의 현황을 나타낸 것이다.

<표2> 시·구 및 군별 경제활동인구와 조사구 수

(단위 : 명, 개)

시·군	남	여	전체	조사구
청주상당구	46,449	38,615	85,064	8
청주흥덕구	77,055	54,506	131,561	14
충주시	53,813	41,745	95,558	11
제천시	41,417	27,383	68,800	4
소계	218,734	162,249	380,983	37
청원군	26,828	22,235	49,063	5
보은군	14,467	13,453	27,920	2
옥천군	22,294	18,431	40,725	3
영동군	22,463	17,353	39,816	3
진천군	12,138	10,197	22,335	1
괴산군	16,244	16,247	32,491	5
음성군	14,985	14,781	29,676	6
단양군	13,030	11,331	24,361	2
소계	142,359	124,028	266,387	27

5.2 시군구 실업자 추정

5.2.1 직접 추정량

직접추정량은 사전에 계산한 동부와 읍면부에 대한 승수(통계청(1996), pp12)와 경활조사 자료만을 이용하여 식(2.1)을 통해서 계산하였다. 또한 분산은 통계청에서 사용하고 있는 연속 차의 분산공식을 적용하여 계산하였으며, 시군구별 실업자 추정값과 분산은 <표3>에 주어졌다. 특기 사항은 진천군의 표본조사구는 1개이므로 분산을 계산할 수 없어서 일반화 분산함수(Wolter(1985), pp210)를 통해서 계산하였다.

5.2.2 합성 추정량

"borrow strength"를 적용하기 위해서 충북을 시 지역과 군 지역으로 크게 2개 그룹으로 구분하고 각 그룹내에서 유사성질 범주의 구분을 성별-연령대별과 성별-교육정도별로 하고

각 셀에 대한 실업률을 추정하였다. 시 지역 내에서는 각 시의 범주별 실업률은 동일하고 군 지역에서 각 군의 범주별 실업률이 동일하다는 조건하에서 소지역별 실업자를 추정한다.

1. 성별-연령대별 구성비를 이용한 시군구별 실업자 추정

연령대별 구분은 15-24세, 25-34세, 35-44세, 45-54세, 55세 이상의 5개 범주로 나누어서 실업률 추정의 정도를 높이고자 하였다.

$$\hat{y}_{i.} = \sum_{j=1} x_{ij}\ r_{.j}^{a} \qquad (5.1)$$

단, i는 시군구를 나타내고, j는 성별-연령대별 분류를 나타낸다. x_{ij}는 i시군구의 j성별-연령대의 경제활동 인구수를 나타내며 주민등록인구수에서 추정하거나 경활조사에서 수집할 수 있으나 여기서는 후자를 이용하였고, $r_{.j}^{a}$는 j성별-연령대의 실업률을 나타내는 것으로, j성별-연령대의 경제활동 인구수에 대한 j성별-연령대의 실업자수를 이용하여

$$r_{.j}^{a} = \frac{y_{.j}}{\sum_{i=1} x_{ij}} \quad , \quad y_{.j} : j\text{성별}-\text{연령대의 실업자수}$$

의 관계로부터 구할 수 있다. 그 결과는 <표4>와 같다.

2. 성별-교육정도별 구성비를 이용한 시군구별 실업자 추정

교육정도별 구분은 초등 학교졸, 중학교졸, 고등 학교졸과 대학교졸(전문학교 포함)로 4개 범주로 나누어 실업률 추정의 정도를 높였다.

$$\hat{y}_{i.} = \sum_{j=1} x_{ij}\ r_{.j}^{b} \qquad (5.2)$$

단, i는 시군구를 나타내고, j는 성별-교육정도별 분류를 나타낸다. x_{ij}는 i시군구의 j성별-교육정도별 경제활동 인구수를 나타내고, $r_{.j}^{b}$는 j성별-교육정도의 실업률을 나타내는 것으로, j교육정도의 경제활동 인구수에 대한 j성별-교육정도의 실업자수를 이용하여

$$r_{.j}^{b} = \frac{y_{.j}}{\sum_{i=1} x_{ij}}, \qquad y_{.j} : j\text{성별}-\text{교육정도의 실업자수}$$

의 관계로부터 계산할 수 있으며, 그 결과는 <표4>와 같다. 식(5.1)과 (5.2)로 주어진 추정량의 분산은 각 범주별 실업률을 상수로 가정하고 x_{ij}의 일차결합의 분산식을 통해서 계산하였다.

5.2.3 복합 추정량

직접 추정량은 표본의 크기에 민감하게 변동하고 합성 추정량은 각 범주들이 모든 시군구에서 동일하다는 전제 조건이 어긋날 경우에는 편향이 커지기 때문에 두 가지 추정량의 문제점을 완화하기 위해서 두 추정값의 가중 평균 형식인 복합 추정량을 생각하게 되었다.

$$\hat{Y}_{i.}^{c} = w_i \ \hat{Y}_{i.}^{d} + (1-w_i) \ \hat{Y}_{i.}^{s} \qquad (5.3)$$

여기에서 $\hat{Y}_{i.}^{d}$는 경활 조사에서 직접 추정한 i 시군구의 실업자 수이고, $\hat{Y}_{i.}^{s}$는 성별-연령대별 분류 또는 성별-교육 정도별 분류에 의해서 합성 추정법으로 추정한 i 시군구의 실업자 수이다. w_i는 $Var(\hat{Y}_{i.}^{d})$와 $mse(\hat{Y}_{i.}^{s})$에 의해서 계산되는 가중값으로 아래와 같이 표현할 수 있다.

$$w_{i(opt)} = \frac{mse(\hat{Y}_{i.}^{s})}{mse(\hat{Y}_{i.}^{s}) + Var(\hat{Y}_{i.}^{d})}$$

여기에서 $mse(\hat{Y}_{i.}^{s})$를 $Var(\hat{Y}_{i.}^{s})$로 대체하여 w_i의 근사값을 다음 식으로 추정하였다.

$$\hat{w}_{i(opt)} = \frac{\widehat{Var}(\hat{Y}_{i.}^{s})}{\widehat{Var}(\hat{Y}_{i.}^{s}) + \widehat{Var}(\hat{Y}_{i.}^{d})}$$

식(5.3)에서 주어진 추정량의 분산은 근사적으로 다음 식으로 계산할 수 있다.

$$\widehat{Var}(\hat{Y}_{i.}^{c}) = \frac{\widehat{Var}(\hat{Y}_{i.}^{s}) \cdot \widehat{Var}(\hat{Y}_{i.}^{d})}{\widehat{Var}(\hat{Y}_{i.}^{s}) + \widehat{Var}(\hat{Y}_{i.}^{d})} \qquad (5.4)$$

8.2.4 추정된 실업자수 조정

성별-연령대별/성별-교육정도별의 각 범주의 실업률을 이용하여 계산된 시 지역의 실업자 합계는 경활 조사 자료에서 직접 추정한 시 지역의 실업자 총수와 같다고 가정하였다. 이를 기준으로 각 시·구의 실업자 추정값을 다음과 같이 조정할 수 있다.

$$\hat{Y}_{i.(A)} = \frac{\hat{Y}_{i.}}{\sum_{i=1}^{4} \hat{Y}_{i.}} \ \hat{Y}^{*}$$

여기에서 $\hat{Y}_{i.}$는 직접추정법과 합성추정법 및 복합추정법으로 계산된 i 시·구의 실업자 수이고, $\hat{Y}^{*}$는 직접 추정한 시지역 실업자 총 수이다.

군 지역의 실업자 총수는 통계청에서 추정한 충청북도의 실업자수에서 시·구 지역의 조정된 실업자 수를 감하여 계산하고, 이 결과를 이용하여 각 군별로 조정된 실업자 추정값

$$\hat{Y}_{i(A)} = \frac{\hat{Y}_{i.}}{\sum_{i=1}^{9} \hat{Y}_{i.}} \ \hat{Y}^{**}$$

로 구한다. 여기에서 $\hat{Y}_{i}$는 직접추정법과 합성추정법 및 복합 추정법에서 추정된 군 지역의 실업자 수이고, $\hat{Y}^{**}$= '충북실업자' − '시·구의실업자합계' 이다.

5.3 추정결과

<표3>의 결과를 살펴보면, 분산 추정에서는 직접추정량의 분산이 가장 크게 나타났다. 합성추정량의 분산은 편향을 생략하였기 때문에 직접추정량보다는 작을 것으로 예상 했지만, 13배에서 200배 정도까지 감소된 것으로 나타났다. 따라서, 표본의 크기가 작을 경우, 합성추정법이 직접추정법에 비해서 효과적임을 알 수 있었으며, 직접추정량에 비해 복합추정량이 보다 안정적인 특성을 보이고 있다. 3가지 추정법 중에서는 복합추정법이 비교적 효과적임이 수치적인 사례로 나타난다.

<표3> 시군구별 실업자 추정값과 분산

구 분	직접 추정량	합성 I	합성 II	복합 I	복합 II
	(통계청)	성별-연령대	성별-교육정도	직접+합성 I	직접+합성 II
청주 상당구	6,847	4,832	5,220	6,837	6,838
	4,149,369	21,413	22,823	21,303	22,699
청주 흥덕구	8,864	8,002	8,522	8,853	8,859
	3,200,521	43,016	50,192	42,446	49,417
충주시	3,234	5,232	5,435	3,265	3,277
	1,758,276	27,833	35,295	27,399	34,600
제천시	3,186	4,065	2,954	3,177	3,157
	1,181,569	15,758	18,807	15,551	18,513
청원군	905	1,488	1,370	911	911
	724,201	7,470	9,010	7,394	8,899
진천군	784	817	790	784	784
	184,041	2,548	2,156	2,513	2,131
괴산군	249	1,021	1,241	309	327
	63,504	5,324	5,447	4,912	5,017
음성군	2,214	1,306	1,454	2,208	2,209
	744,769	4,893	5,066	4,861	5,032
보은군	915	573	624	915	914
	763,876	1,542	1,575	1,539	1,572
옥천군	1,207	834	831	1,192	1,189
	81,225	3,400	4,023	3,263	3,833
영동군	593	931	673	616	599
	41,616	3,044	3,103	2,837	2,888
단양군	744	641	628	676	678
	57,121	2,239	2,469	2,155	2,367
합 계	29,742	29,742	29,742	29,742	29,742

<표4> 충북지역의 특성별 실업률

구 분		남		여	
		시지역	군지역	시지역	군지역
연령대별	15-24	0.1244	0.0551	0.0515	0.1163
	25-34	0.0895	0.0817	0.0454	0.0177
	35-44	0.0668	0.0283	0.0545	0.0081
	45-54	0.0201	0.0325	0.0387	0.0173
	55세이상	0.0334	0.0094	0.0240	0.0089
교육 정도별	초졸	0.0427	0.0060	0.0161	0.0055
	중졸	0.0564	0.0171	0.0428	0.0219
	고졸	0.0684	0.0542	0.0510	0.0473
	대졸	0.0697	0.0723	0.0506	0.0000

Ⅵ. 결 언

소지역 추정에 관한 연구는 최근 몇몇 학자들에 의해 꾸준히 진행되고 있으며, 앞서 소개되었던 소지역 추정법을 근간으로 하여 몇가지 확장된 이론들이 소개되고 있다.

시계열 모형이 소지역 추정 문제에 적용되고 있다. 시간 t에서 i소지역에 대한 특성 모수를 θ_{it}라 하고, $\hat{\theta}_{it}$를 θ_{it}의 직접추정량이라 하자. 표본모형은 θ_{it}가 주어진 상태에서 $\hat{\boldsymbol{\theta}}_i=(\hat{\theta}_{i1},\cdots,\hat{\theta}_{iT})^T$가 평균이 $\boldsymbol{\theta}_i=(\theta_{i1},\cdots,\theta_{iT})^T$이고 기지인 공분산 $\boldsymbol{\psi}_i$를 갖는다고 가정할 때 다음과 같은 연결모형을 고려한다.

$$\theta_{it}=\boldsymbol{x}_{it}^T\boldsymbol{\beta}+v_i+u_{it}$$

여기에서 v_i는 평균이 0, 분산이 σ_v^2인 서로 독립인 정규분포를 따르는 것으로 가정되며, u_{it}는 일계 자기회귀 모형 $u_{it}=\rho u_{i,t-1}+\varepsilon_{it}$ ($|\rho|<1$), 또는 랜덤워크 모형 $u_{it}=u_{i,t-1}+\varepsilon_{it}$을 따르는 것으로 가정된다. e_{it}는 v_i와는 독립이며 평균이 0, 분산이 σ^2 인 정규분포를 따른다. 이러한 연결 모형이 계량경제학 분야에서 폭넓게 연구되고 있다.

Moura and Holt(1999)는 기초적인 단위 수준 모형(unit-level model)을 일반화하여 단위 수

준(unit-level)과 지역 수준(area-level) 공변량들을 하나의 모형으로 통합한 2 수준(two-level) 모형을 이용하여 소지역 추정을 시도하였다. Datt, Day and Basawa(1999)는 기초적인 단위 수준(unit-level) 모형을 다변량의 경우로 확장하여 다변량 내포오차회귀모형을 이끌어 냈다.

최근에 들어서는 이진 자료(binary data, $y_{ij}=0$ or 1)에 대해서 로지스틱 선형혼합모형을 적합시키는 문제들이 연구되고 있다. 표본모형은 주어진 θ_{ij}에 대해서 y_{ij}가 모수 θ_{ij}를 갖는 독립인 베르누이 변수들임을 가정한다. 연결모형은 v_i를 갖는 일종의 로지스틱 회귀모형인 $\frac{\theta_{ij}}{1-\theta_{ij}} = \boldsymbol{x}_{ij}^{T} + v_i$ 이며, v_i는 서로 독립이고 동일한 분포를 갖는 정규분포이며, 평균이 0, 공통 분산이 σ_v^2임을 가정한다. 즉, 연속형의 자료뿐만 아니라 이진 자료를 하나의 모형으로 통합하여 소지역 추정 문제를 해결하는 연구들도 진행 중에 있다.

〈참 고 문 헌〉

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

[2] 통계청(1996), 「표본개편 연구회 연구 보고」

[3] Cowles,M.K. and Carlin,B.P.(1996) Markov Chain Monte Carlo convergence diagonstics: a comparative review. *Journal of the American Statistical Association,* 91, 883-904

[4] Datta, G.S., Day, B. and Basawa, I. (1999) Empirical best linear unbiased and empirical Bayes prediction in multivariate small area estimation. *Journal of Statistical Planning and Inference,* 75, 269-279

[5] Erickson, E.P.(1974) A regression method for estimating populations of local areas. *Journal of the American Statistical Association,* 69, 867-875

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

[7] Gonzalez, M.E. (1973) Use and evaluation of synthetic estimates, *Proceedings of the Social Statistics Section,* American Statistical Association, pp. 33-36

[8] Hobert, J.P. and Casella, G. (1996) The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association,* 91, 1461-1479

[9] Jiang, J. (1996) REML esitmation: asymptotic behaviour and related topics. *Annals of Statistics,* 24, 255-286

[10] Lahiri, P.A. and Rao, J.N.K. (1995) Robust estimation of mean squares error of small area estimators. *Journal of the Americal Statistical Association,* 82, 758-766

[11] Leslie Kish(1965), *Survey Sampling*, John Wiley & Sons Inc., New York

[12] Morris H. Hansen, William N. Hurwitz, and William G. Madow(1993), *Sampling Survey Methods and Theory Vol I &II*, John Wiley & Sons Inc., New York

[13] Moura, F. and Holt, D. (1999) Small area estimation using multilevel models. *Survey Methodology,* 25, 73-80

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

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

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

[17] William G. Cochran(1977), *Sampling Techniques 3rd ed.,* John Wiley & Sons Inc., New York