소지역통계 추정법
1차 연구결과 자료

소 지 역 통 계 추 정 법

- 충북의 시·군·구별 실업통계 추정결과를 중심으로 -

2001. 1

통 계 기 획 국
조 사 관 리 과

머 리 말

이 보고서는 시·군·구 단위등의 소지역통계 수요에 부응하기 위하여 사계 전문가인 공군사관학교 이계오교수와 공동으로 수행한 소지역통계 추정기법에 대한 1차 연구 결과물입니다.

또한, 이 보고서의 내용은 각종 소지역통계 추정 기법과 미국, 캐나다 등의 적용사례를 비교 설명하였으며, 이를 토대로 우리나라의 실정에 맞는 소지역통계 추정기법 모형 개발을 위하여 일부 시험지역(충북)의 시·군·구별 실업통계의 추정 결과를 수록하고 있습니다.

이러한 연구 작업을 기초로 금년중에는 각종 관련 행정자료와 매월 경상조사인 경제활동인구조사 결과를 이용한 각 지역별 추정량 모델개발 연구를 지속해 나갈 계획입니다.

이와 같은 연구작업 결과는 2002년 가구부문 표본개편 및 설계작업시에 반영되어 날로 늘어나는 소지역통계 수요에 부응할 수 있도록 활용되어질 것입니다.

끝으로, 소지역통계 추정과 관련하여 관심 있는 여러분께서 좋은 의견을 주시면 추정 모델 개발 과정에 적극 수렴하여 반영해 나가도록 하겠습니다.

2001년 1월

조사관리과장 **김 상 식**

목 차

1. 서 언

지식기반 정보화 사회로 발전과 지방자치제도의 활성화로 좀더 정확하고 시의적절한 통계생산의 필요성이 높아지고 있을뿐아니라 지금까지 생산하지 않았던 소지역단위(전국단위 통계에서는 시도단위 통계 또는 시도단위통계가 생산되고있는 통계는시군구단위통계를 의미함)통계에 대한 요구와 필요성이 대두되었으나 우리나라에서는 구체적인 연구가 미진하여 통계선진국들의 연구사례를 살펴봄으로써 우리나라에서 소지역 통계작성의 가능성을 제시하여 연구기반을 마련하고자 한다.

미국이나 캐나다 등의 통계 선진국에서는 센서스의 중간 년도에 해당되는 해에 주 또는 county의 인구 추정과 노동력 통계를 생산하기 위해 전국적인 대규모의 표본조사를 실시하고 있다. 표본 설계 당시에는 대지역 단위의 통계를 생산할 목적으로 표본조사를 시행하였으나 지방자치정부의 요청이나 중앙 정부의 예산 배정을 위해서 좁은 지역단위의 통계를 생산할 필요가 있다. 이 경우에는 신뢰성있는 통계생산을 위하여 직접 조사된 자료뿐만 아니라 행정보고 자료 또는 센서스 자료 구조의 특성을 이용하여 신뢰성을 높일 수 있는 간접추계법인 소지역 추정법을 활용하고 있다.

표본설계 당시에는 통계작성단위의 규모가 컸으나 자료 수집 후에 통계작성단위의 세부단위에 대한 통계를 생산할 필요가 있을 경우에는 세부단위에서 조사된 자료의 수가 적거나 아예 없을 수 있으므로 세부단위에 대한 추정값은 추정오차가 클뿐만 아니라 세부단위들 간의 표본수의 차이 때문에 추정오차의 크기도 천차만별이 될 것이다. 예를 들면 경활조사에서 시도별로 실업, 취업 등의 통계를 생산하고자 한다면 각 시군구에 할당된 표본조사구 수의 분포가 불균등하므로 시군구 단위별 실업통계의 신뢰성이 크게 저하될 수 있으며 또 다른 경우로는 각 시도에서 인구사회학적 분류인 성별-연령대별로 구분하여 각 Cell에서 실업률이나 실업자에 대한 통계를 구하고자 할 경우에도 각 Cell에 분포된 표본수가 불균형적일 뿐만 아니라 적기 때문에 추정오차가 커질 것을 생각할 수 있다.

세부단위에 배정된 표본수가 적을 경우에는 특성이 유사한 인근의 세부단위들을

결합하여 세부단위들의 그룹을 만들고 그 그룹 내에서 세부단위들이 동일한 특성을 갖는다는 조건에서 조사된 자료를 이용하여 세부단위의 통계를 작성하거나, 행정업무자료 또는 센서스 등 다른 통계조사 정보를 이용하여 세부단위의 통계를 작성하는 기법을 소지역 추정법(Small Area Estimation)이라고 한다.

우리나라에서는 취업, 실업 등과 같은 경제적 특성을 조사하여 국가의 고용정책 입안과 평가에 필요한 기초자료를 수집하기 위해서 매월 3만 표본가구내에 거주하는 만 15세 이상인 사람들을 대상으로 경제활동 전반에 걸쳐서 조사하고 있다. 매월 15일을 포함하는 조사대상 주간 중에 표본가구에 거주하는 사람들의 취업, 실업 및 비경제활동인구 관련사항을 방문면접이나 컴퓨터 면접방식으로 조사한 후 조사된 자료를 이용하여 매 익월 말경에 전국단위와 각 시도단위로 경제활동 관련 통계를 발표하고 있다.

1995년부터 시작된 지방자치제도와 1997년의 IMF사태로 실업자 구제대책의 효과적인 시행을 위해서 각 시도내의 시군구별 실업관련 통계작성이 요구되었을 뿐만 아니라 지방자치행정의 발전적인 정착을 위해서는 다양한 분야의 통계도 시군구 단위까지 작성해야 한다는 인식이 높아지고 있지만 현행의 통계작성 방법으로 모든 분야의 통계를 시군구 단위까지 작성하기 위해서는 현재의 몇배의 시간, 예산과 인원이 소요될 것이므로 실현 가능성이 희박하다. 그러나 통계 선진국인 미국이나 캐나다에서 소지역의 통계작성에 이용되고 있는 소지역 추정법을 우리나라에서도 적용할 수 있다면 현재의 예산과 인원규모로도 시군구 단위의 통계를 작성할 수 있음을 살펴보고자 한다.

본 연구에서는 소지역 추정법의 개념을 정리하고, 시군구 실업자 추정에 이용될 수 있는 추정법들을 요약하며 미국의 소지역 추정법을 이용한 소지역의 실업통계 작성 과정을 살펴본다. 또한 영국과 캐나다 등에서 적용되는 소지역 추정법을 살펴본다. 마지막으로 소지역 추정법을 이용한 시군구의 실업자 추정에 대한 사례를 경활 조사 자료를 통해서 제시할 것이다.

2. 직접 추정법(Direct Estimation)

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 i시군구에서 조사한 조사구들의 집합이다.

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

3. 간접 추정법(Indirect Estimation)

3.1 인구통계적 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 생존률법(Vital Rates Method: VR Method)

VR법은 출생과 사망에 관련된 자료를 이용하여 인구의 변동률보다는 장후 변수의 영향만을 분석하여 활용한다. 가장 최근에 센서스를 실시한 해를 기준 년도로 하고, 기준해로부터 t년 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.

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

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

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 t년 후의 인구수는 다음 식에 의해서 추정할 수 있다.

$$p_t = \frac{1}{2}\left(\frac{b_t}{\gamma_{bt}} + \frac{d_t}{\gamma_{dt}}\right) \tag{3.2}$$

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 인구 수를 추정하기위해 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.

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

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), 차분 상관계수(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 소지역 i 를 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 m개의 소지역 중 k개의 소지역에 대하여 선형회귀식을 적합시켜 회귀계수들

을 추정한 후, Y_i 의 추정값으로 다음의 표본 회귀추정량을 이용한다.

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

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

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

표본 회귀추정량은 표본으로부터 직접 추정된 값이 아니라 다중선형회귀를 거쳐 얻어진 보정된 추정량이며, 표본 회귀법은 이를 이용하여 소지역의 인구를 추정하는 방법이다. 그러나 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. 우리나라의 경활조사의 예로 살펴보면 조사단위는 동부와 읍면부로 나누어져 광역단체별로 층화추출되므로 시 지역과 읍면 지역으로 그룹을 나눈다면 각 그룹내에서는 연령대별 구조나 교육정도별 구조의 특성이 거의 유사할 것으로 판단할 수 있고, 이러한 경우 시군구 실업자 추정에 합성추정법의 이용도 가능할 것으로 생각된다.

대영역을 I개 소지역으로 분할하며 또한 대영역을 특성 기준에 따라 J개의 범

주로 분류한다면 i소지역의 추정값은 다음 식으로 구할 수 있다.

$$\hat{Y}_{i\cdot} = \sum_{j} p_{ij}\ \hat{Y}_{\cdot j} \tag{3.6}$$

단, p_{ij}는 i번째 소지역의 j범주에 대한 가중값이며 센서스나 행정자료에서 구해진다. $Y_{\cdot j}$는 대영역에서 j범주에 대한 표본에서 구한 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 합성 추정량은 다음과 같이 나타낼 수 있다.

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

만일에 $\hat{Y}_{d\cdot j}$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.

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

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 되고, 이때 $\hat{Y}_{i.}^{s}$의 편향(Bias)의 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 불편추정량이 아래와 같이 주어질 수 있다.

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

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.

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

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

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

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

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

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

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

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

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 생존불법(VR Method), 합성추정법 또는 복합 추정법 중 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 i소지역의 *추정법으로 추정한 것이다.

4. 모형 기반 추정법(Model-Based Estimation)

4.1 기본적인 지역 수준 모형(Basic Area-level Model)

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 있으며, 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 최량선형 불편예측(EBLUP) 추정량, 경험적 베이즈(EB) 추정량, 계층적 베이즈(HB) 추정량에 대해 설명하기로 한다. 지역 수준 모형은 기본적으로 두가지의 성분들로 이루어진다. 즉, 소지역에 대한 직접추정량 $\hat{\theta}_i$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 $\hat{\theta}_i = g(\hat{\overline{Y}}_i)$은 모집단의 값 $\theta_i = g(\overline{Y}_i)$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 표본오차 e_i 는 서로 독립이며, 평균이 0, 분산이 ψ_i 임이 가정되며, 보

통 ϕ_i 는 알려진 것으로 가정한다.

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

$$\theta_i = z_{i1}\beta_i + z_{i2}\beta_2 + \cdots + z_{ip}\beta_p + v_i$$

$$= \boldsymbol{z}_i^T \boldsymbol{\beta} + v_i \tag{4.2}$$

여기에서 모형오차 v_i 는 서로 독립이며, 평균이 0, 분산 σ_v^2을 갖고, 표본오차 e_i와는 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^2 은 소지역들의 동질성을 나타내는 측도이다.

4.2 경험적 최량선형불편예측(EBLUP) 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, MSE추정을 위해 정규분포를 가정하기도 한다. 또한, EBLUP 추정량과 EB 추정량은 e_i와 v_i를 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).

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

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.

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

(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 회귀합성추정량으로 주어질 수 있다.

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

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

단, $g_{1i}(\sigma_v^2) = \gamma_i \psi_i$ 이고, $g_{2i}(\sigma_v^2) = \sigma_v^2 (1-\gamma_i)^2 \; z_i^T (\sum_i \gamma_i z_i \; z_i^T)^{-1} z_i$ 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 $O(m^{-1})$의 형식 유계인 항이며, 이로부터 BLUP 추정량의 MSE 값은 γ_i 나 모형분산 σ_v^2이 표본분산 ψ_i에 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 미지인 값이므로 적절한 $\hat{\sigma}_v^2$을 추정하여 EBLUP 추정량 $\hat{\theta}_i = \hat{\theta}_i(\hat{\sigma}_v^2)$을 산출한다. 이때 소지역의 평균 $\overline{Y_i}$ 의 추정량은 $g^{-1}(\hat{\theta}_i)$ 로, σ_v^2의 추정량은 $\hat{\sigma}_v^2 = \max(\tilde{\sigma}_v^2, 0)$로 주어진다. 여기에서 $\tilde{\sigma}_v^2$은 다음 식을 만족한다.

$$(m-p)\;\tilde{\sigma}_v^2 = \sum_i (\hat{\theta}_i - \; z_i^T \; \beta^*)^2 - \sum_i \psi_i h_{ii} \qquad (4.7)$$

(4.7)식에서 $h_{ii} = \; z_i^T (\sum_i \; z_i \; z_i^T)^{-1} z_i$ 이고, β^*는 β의 OLS(ordinary least squares) 추정량이다. 한편, $\tilde{\sigma}_v^2$은 다음과 같은 비선형 방정식의 반복적인 해로써 구할 수도 있다.

$$a(\sigma_v^2) = \sum_i \{\hat{\theta}_i - \; z_i^T \tilde{\beta}(\sigma_v^2)\}^2 / (\sigma_v^2 + \psi_i) = m-p \qquad (4.8)$$

여기에서 $\tilde{\beta}(\sigma_v^2)$은 (4.5)식에 주어졌고, (4.8)식의 가운데 항은 가중잔차제곱합, $m-p$ 는 가중잔차제곱합과 관계가 되는 자유도이다. 만약 $\hat{\sigma}_v^2 = 0$ 이면 EBLUP 추정량 $\hat{\theta}_i$ 는 회귀합성추정량 $z_i^T\hat{\beta}$ 로 축약된다. 단, $\hat{\beta} = \tilde{\beta}(\hat{\sigma}_v^2)$이며, 식(4.5)에서 σ_v^2 대신에 $\hat{\sigma}_v^2$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, $\hat{\theta}_i$는 평균이 $z_i^T\beta$ 이고 분산이 $\sigma_v^2 + \psi_i$ 인 서로 독립인 정규분포를 따르게 된다. 이러한 분포에 대한 가정하에서 계산된 β 와 σ_v^2의 최대우도추정량을 제한최대우도추정량(REML)이라 하며, 선형혼합모형하에서도 근사적으로 유효하다. 따라서 $\hat{\theta}_i$ 의 BLUP 추정량을 산출 시 σ_v^2의 REML 추정량을 이용하여도 근사적으로 타당하다.

4.3 경험적 베이즈(EB) 방법

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

$$\hat{\theta}_i^B(\beta, \sigma_v^2) = E(\theta_i|\hat{\theta}_i, \beta, \sigma_v^2) = \gamma_i\hat{\theta}_i + (1-\gamma_i)\, z_i^T\beta \qquad (4.9)$$

(4.9)식의 베이즈 추정량은 선형성 또는 불편성을 만족하지는 않는다. 여기에서 모수 β 와 σ_v^2을 제한최대우도(REML) 추정량으로 대체하여 다음과 같은 θ_i 의 경

험적 베이즈(EB) 추정량을 얻는다.

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

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

EBLUP 추정량 $\hat{\theta}_i = \hat{\theta}_i(\hat{\sigma}_v^2)$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 ***MSE***의 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 $\hat{\theta}_i$ 의 ***MSE*** 추정량을 제안하였다. Prasad and Rao(1990)가 제안한 ***MSE*** 추정량은 (4.7)식의 σ_v^2의 적률추정량을 사용하였을 경우 다음과 같이 주어진다.

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

단, $g_{1i}(\sigma_v^2) = \gamma_i \psi_i$, $g_{2i}(\sigma_v^2) = \sigma_v^2(1-\gamma_i)^2 \; z_i^T(\sum_i \gamma_i z_i \; z_i^T)^{-1} z_i$,

$g_{3i}(\sigma_v^2) = \{\psi_i^2/(\sigma_v^2 + \psi_i)^3\} h(\sigma_v^2)$, $h(\sigma_v^2) = 2m^{-2}\sum_i(\sigma_v^2 + \psi_i)^2$로 주어진다.

최근들어 Jing, Lahiri and Wan(1999)는 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 $\hat{\theta}_i^{EB} = k(\hat{\theta}_i, \hat{\varphi})$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 $(\hat{\theta}_l, z_l)$을 제외한 φ 의 추정량 $\hat{\varphi}(l)$을 계산한다. 이때의 EB 추정량을 $\hat{\theta}_i^{EB}(l) = k(\hat{\theta}_i, \hat{\varphi}(l))$로 나타내자.

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

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

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

4.4 계층적 베이즈(HB) 방법

계층적 베이즈(HB) 방법을 이용한 추론은 비교적 추론의 정확도가 높고, 복잡한 유형의 문제들에서도 최근에 개발된 MCMC(Markov Chain Monte Carlo)방법을 이용하여 해결할 수 있다. 깁스 샘플러도 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 들에 관한 추론은 주변 사후분포에 의

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

먼저 σ_v^2 이 기지인 상태를 가정하고 $\boldsymbol{\beta}$ 에 관한 사전분포를 배정하기로 한다. $\boldsymbol{\beta}$의 사전분포가 상수에 비례하고(i.e improper prior), v_i 와 e_i 가 정규분포를 따른다고 가정한다면, 이때 사후평균 $E(\theta_i|\ \hat{\boldsymbol{\theta}},\ \sigma_v^2)$은 (4.4)식의 BLUP 추정량 $\hat{\theta}_i(\ \sigma_v^2)$과 동일하다. 더욱이 사후분산 $V(\theta_i|\ \hat{\boldsymbol{\theta}},\ \sigma_v^2)$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 $\boldsymbol{\beta}$ 뿐만 아니라 σ_v^2에 관한 사전분포를 고려해야 하며, 또한 서로 독립임을 가정하여 주변사후분포 $f(\ \sigma_v^2|\ \hat{\boldsymbol{\theta}})$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 $G(a,\ b)$와 같이 배정한다(여기에서 $G(a,\ b)$: $f(\tau_v) \propto \exp(-a\tau_v)\ \tau_v^{b-1}$). 주변사후분포 $f(\ \sigma_v^2|\ \hat{\boldsymbol{\theta}})$를 이용한 HB 추정량 $E(\theta_i|\ \hat{\boldsymbol{\theta}})$은 다음 식과 같이 주어진다.

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

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

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

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

깁스 샘플링은 위의 (4.12)와 (4.13)을 결정하기위해 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) $\theta_i \mid \boldsymbol{\beta}, \sigma_v^2, \hat{\boldsymbol{\theta}} \sim N(\hat{\theta}_i^B(\boldsymbol{\beta}, \sigma_v^2), g_{1i}(\sigma_v^2) = \gamma_i \psi_i)$

(iii) $\tau_v = \sigma_v^{-2} \mid \boldsymbol{\beta}, \boldsymbol{\theta}, \hat{\boldsymbol{\theta}} \sim G(\tilde{a}, \tilde{b})$, 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 위의 (i)로부터 $\boldsymbol{\beta}^{(1)}$ 계산

(b) $\boldsymbol{\beta} = \boldsymbol{\beta}^{(1)}$, $\sigma_v^2 = \sigma_v^{2(0)}$를 이용하여 위의 (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 위의 (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 $\theta_i^{(0)} = \hat{\theta}_i^{EB}$, $\sigma_v^{2(0)} = \sigma_v^2$의 REML 추정량을 사용한다.

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

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

$$= \frac{1}{J}\sum_j \hat{\theta}_i(j) = \hat{\theta}_i(\cdot), \tag{4.14}$$

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

5. 소지역 추정 적용 사례

5.1 Multi-Level 모형을 이용한 소지역 추정 사례(I)

Multilevel 모형은 소지역 내의 변동과 소지역 간의 변동을 함께 고려하여 소지역 추정의 정확성을 높이기 위하여 제안되었던 방법이며, 조사자료를 이용하여 가정한 Multilevel 모형의 모수를 추정하고 구하고자 하는 소지역의 특성값을 예측한다.

5.1.1 Multilevel Model Framework

다음과 같이 단위수준과 지역수준 공변량을 하나의 모형으로 통합한 다수준 모형을 고려한다.

$$Y_i = X_i \beta_i + \varepsilon_i \ , \tag{5.1}$$

$$\beta_i = Z_i \gamma + \nu_i \quad , \quad \text{단, } i = 1, 2, \cdots, m$$

여기에서 $Y_i = i$번째 소지역에서 길이가 n_i인 벡터, $X_i = i$번째 소지역에서 설명변수들의 행렬($(p+1) \times n_i$ 행렬), $Z_i =$ 소지역 변수들의 Design Matrix, $\gamma =$ 길이가 q인 벡터, $\beta_i = (\beta_{i0}, \beta_{i1}, \cdots, \beta_{ip})^t$는 i번째 소지역에서 길이가 $p+1$인 회귀계수의 벡터, $\nu_i = (\nu_{i0}, \nu_{i1}, \cdots, \nu_{ip})^t$는 i번째 소지역에서 길이가 $p+1$인 랜덤효과의 벡터를 나타내며, 오차항 $\varepsilon_i \overset{ind}{\sim} N(0, \sigma^2 I)$, $\nu_i \overset{ind}{\sim} N(0, \Omega)$ 를 가정하며, ε_i와 ν_i는 서로독립임을 가정한다.

i번째 소지역에 대해서 위의 모형을 구체적으로 명시하면 다음과 같이 표현될 수 있다.

$$Y = X\beta + \varepsilon$$

$$= XZ_\gamma + X\nu + \varepsilon \ , \quad \text{단, } Z_\gamma = Z\gamma$$

$$;\quad \begin{pmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ Y_i \\ \cdot \\ \cdot \\ Y_m \end{pmatrix} = \begin{pmatrix} X_1 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & X_2 & 0 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & X_i & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdot & \cdot & \cdot & X_m \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \beta_i \\ \cdot \\ \cdot \\ \beta_m \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \varepsilon_i \\ \cdot \\ \cdot \\ \varepsilon_m \end{pmatrix}$$

$$Y \qquad\qquad X \qquad\qquad \beta \qquad \varepsilon$$

여기에서 $Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \cdot \\ \cdot \\ \cdot \\ Y_{in_i} \end{pmatrix}$, $\quad X_i = \begin{pmatrix} 1 & x_{i11} & x_{i21} & \cdot & \cdot & x_{ip1} \\ 1 & x_{i12} & x_{i22} & \cdot & \cdot & x_{ip2} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_{i1n_i} & x_{i2n_i} & \cdot & \cdot & x_{ipn_i} \end{pmatrix}$

$$\beta_i = \begin{pmatrix} \beta_{i0} \\ \beta_{i1} \\ \cdot \\ \cdot \\ \cdot \\ \beta_{ip} \end{pmatrix} \; (= Z_i \gamma + \nu_i \;), \qquad \varepsilon_i = \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_{in_i} \end{pmatrix}$$

즉, 위의 모형은 회귀계수에 대한 랜덤효과를 모형에 포함시켜 단위수준과 지역수준 공변량을 한 모형으로 통합한 모형구조를 갖고 있다.

5.1.2 고정 모수 γ , 분산성분 모수 $\theta = ([\,\mathrm{Vech}\,(\Omega)]^t, \sigma^2)^t$ 의 추정

γ 와 θ 의 추정은 정규조건 하에서 ML 방법과 REML(Restricted ML) 방법을 이용하여 추정하는 방법, IGLS(Iterative Generalized Least Squares) 절차를 이용하여 추정하는 방법, RIGLS(Restricted IGLS) 절차를 이용하여 추정하는 방법이 있으며, 이 경우 IGLS 추정량은 일치추정량이 되고, 정규성 가정하에서 MLE와 동치이며, RIGLS 추정 절차는 분산성분 모수들의 불편추정량을 제공하며 정규성 가

정하에서 REML 추정량과 동치이다.

IGLS 절차를 이용하여 γ 와 θ 의 추정하는 방법은 다음과 같다.

i) $\theta = ([\ \mathrm{Vech}(\Omega)]^t , \sigma^2)^t$ 의 초기값을 setting하여 $\tilde{\gamma}$ 를 다음식으로 추정한다.

$$\tilde{\gamma} = (Z^t X^t V^{-1} X Z)^{-1} (Z^t X^t V^{-1} Y)$$

$$= (\sum_{i=1}^{m} Z_i^t X_i^t V_i^{-1} X_i Z_i)^{-1} (\sum_{i=1}^{m} Z_i^t X_i^t V_i^{-1} Y_i) \quad (5.2)$$

여기에서 $V_i = \sigma^2 I + X^t \Omega X$ 는 Y_i 의 공분산 행렬이며, $Z_i =$ design matrix , $V = Diag(V_1 , V_2 , \cdots , V_m) = Diag(V(Y_1) , V(Y_2) , \cdots , V(Y_m))$를 나타낸다.

ii) 위에서 추정된 $\tilde{\gamma}$ 값과, θ 의 초기값을 setting하여 θ 의 개선된 추정값 $\tilde{\theta_a}$ 를 다음식으로 추정한다.

$$\tilde{\theta_a} = Cov(\tilde{\theta_a}) (\frac{\partial\ \mathrm{Vech}(V)}{\partial \theta})^t (\frac{1}{2} V^{-1} \otimes V^{-1})\ \mathrm{Vech}(\hat{Y}\ \hat{Y}^t), \quad (5.3)$$

단, $\hat{Y} = Y - XZ_{\tilde{\gamma}}$, $Cov(\tilde{\theta_a}) = \left\{ (\frac{\partial\ \mathrm{Vech}(V)}{\partial \theta})^t (\frac{1}{2} V^{-1} \otimes V^{-1}) (\frac{\partial\ \mathrm{Vech}(V)}{\partial \theta}) \right\}^{-1}$ 를 나타낸다.

모형 (5.1)를 가정하였을 때 γ 의 GLSE(Generalized Least Squares Estimator) $\tilde{\gamma}$ 는 (5.2)식과 같이 유도할 수 있으나, V 가 미지이므로 γ 는 (5.2)식을 이용하여 직접적으로 추정할 수는 없다. 따라서 θ 의 초기값을 setting하여 반복 알고리즘을 적용하여 γ 를 추정한다.

추정된 $\tilde{\gamma}$ 를 이용하여 V 를 추정(θ 를 추정)함에 있어 다음의 사실을 이용하여, 먼저 γ 가 기지인 상태에서 V 를 추정한다.

i) $\mathrm{Vech}(V) = E[\ \mathrm{Vech}\{(Y - XZ_\gamma)(Y - XZ_\gamma)^t\}]$

$= E[\ Y^*]$

; Y^* 는 $\mathrm{Vech}(V)$의 불편추정량

ii) $\mathrm{Vech}(V)$는 $\theta = (\ [\ \mathrm{Vech}(\Omega)]^t\ ,\ \sigma^2\ )^t$ 의 선형함수

위의 i), ii)를 이용하여 다음과 같은 선형함수를 생각하자.

$$Y^* = F\theta + \zeta\ , \tag{5.4}$$

단, $F = \frac{\partial\ \mathrm{Vech}(V)}{\partial\ \theta}$, $\zeta =$ 평균이 $\mathbf{0}$인 확률변수, $V_\zeta = 2\phi_n(V \otimes V)\ \phi_n{}^t$, 여기에서 ϕ_n 은 $\mathrm{Vech}(A)$에서 $\mathrm{Vech}(A)$로 가는 임의의 선형변환이고, A 는 $\mathrm{Vech}(A) = \phi_n\ \mathrm{Vech}(A)$ 를 만족하는 $n \times n$ 행렬이다.

위 (5.4)식의 Y^* 에서 V_ζ 를 기지이며, 정칙행렬로 가정하여 θ 의 GLSE를 (5.3)식과 같이 추정한다. 이때 (5.3)식의 $\tilde{\theta}_a$ 는 θ 의 초기값을 setting하여 반복적으로 계산 추정된다.

RIGLS 절차를 이용하여 γ 와 θ 를 추정하는 방법은 다음과 같다.

i) $\theta = (\ [\ \mathrm{Vech}(\Omega)]^t\ ,\ \sigma^2\ )^t$ 의 초기값을 setting하여 $\tilde{\gamma}$ ($= \hat{\gamma}$)를 추정한다(IGLS 절차에서 $\tilde{\gamma}$ 추정과 동일함).

ii) i)에서 추정된 $\hat{\gamma}$ 값과, θ 의 초기값을 setting하여 $\tilde{\theta}_a$ 를 계산하는데, 여

기에서는 IGLS의 (5.3)식의 항들 중 $\hat{Y}\ \hat{Y}^t = (Y - XZ_{\hat{\gamma}})(Y - XZ_{\hat{\gamma}})^t$ 대신에 $\hat{Y}\ \hat{Y}^t = (Y - XZ_{\hat{\gamma}})(Y - XZ_{\hat{\gamma}})^t + XZ(Z^t X^t V^{-1} XZ)^{-1} Z^t X^t$를 이용한다.

V 가 기지인 상태에서 GLS방법으로 γ 가 추정된다면 다음의 식

$E\{(Y - XZ_{\hat{\gamma}})(Y - XZ_{\hat{\gamma}})^t\} = V - XZ(Z^t X^t V^{-1} XZ)^{-1} Z^t X^t$ 이 성립한다. V 의 근사적인 불편추정량 (즉, θ 의 불편추정량)을 얻기위한 방법으로 반복계산 시 $\hat{Y}\ \hat{Y}^t = (Y - XZ_{\hat{\gamma}})(Y - XZ_{\hat{\gamma}})^t + XZ(Z^t X^t V^{-1} XZ)^{-1} Z^t X^t$ 를 이용한다.

5.1.3 소지역 평균의 추정량

(1) μ_i 의 EBLUP 추정량

소지역의 특성치를 추정하기 위하여 먼저 다음과 같은 모형을 가정하기로 한다.

$$Y_i = X_i \beta_i + \epsilon_i \ ,$$

$$\beta_i = Z_i \gamma + \nu_i \quad , \quad i = 1, 2, \cdots, m \text{ (소지역 수)}$$

여기에서 N_i 는 i 번째 소지역의 모집단 크기를 나타낸다.

위의 모형에서 i 번째 소지역에 대한 평균의 기댓값은 다음의 (5.5)식과 같이 표현할 수 있다.

$$\mu_i = \overline{X_i}^t \beta_i$$

$$= \overline{X}_i^t (Z_i \gamma + \nu_i)$$

$$= \overline{X}_i^t Z_i \gamma + \overline{X}_i^t \nu_i , \qquad (5.5)$$

여기에서 $\overline{X}_i = (1, \bar{x}_{i1}, \bar{x}_{i2}, \cdots, \bar{x}_{ip})^t$ 는 i 번째 소지역에 대한 모집단 평균벡터를 나타낸다.

μ_i 의 EBLUP 추정량 $\hat{\mu}_i$ 은 다음의 (5.6)식과 같이 표현된다.

$$\hat{\mu}_i = \overline{X}_i^t Z_i \hat{\gamma} + \overline{X}_i^t \hat{\nu}_i , \qquad (5.6)$$

단, $\hat{\nu}_i = \hat{\Omega} X_i^t \hat{V}_i^{-1} (Y_i - X_i Z_i \hat{\gamma})$, $\hat{V}_i^{-1} = \hat{\sigma}^{-2} I - \hat{\sigma}^{-4} X_i \hat{\Omega} \hat{G}_i^{-1} X_i^t$, $\hat{G}_i^{-1} = (I + \hat{\sigma}^{-2} X_i^t X_i \hat{\Omega})^{-1}$이며, 여기에서 γ 와 θ 의 추정량 $\hat{\gamma}$ 와 $\hat{\theta}$ 는 앞서 계산된 RIGLS 추정량이 사용된다.

절편항만 랜덤이고, β 의 나머지 항들은 고정계수들인 Battese et al.(1981, 1988)이 제안한 μ_i 의 EBLUP 추정량 $\hat{\mu}_{i(RI)}$ 은 다음의 (5.7)식과 같다.

$$\hat{\mu}_{i(RI)} = \overline{X}_i^t \hat{\beta} + \hat{\nu}_{i0} \qquad (5.7)$$

(2) $\hat{\mu}_i$의 *MSE* 근사와 추정식

다음의 사실을 이용하여 $\hat{\mu}_i$ 의 MSE 근사식을 유도한다.

i) μ_i 의 REMLE 는 translation invariant (Kackar & Harville(1984)).

ii) μ_i 의 RIGLSE 는 정규성 가정 하에서 REMLE와 동치(Goldstein(1989))

위의 사실을 이용한다면, μ_i 의 RIGLSE 인 $\hat{\mu}_i$ 도 translation invariant 이므로

다음 식이 성립하게 된다.

$$MSE(\widehat{\mu_i}) = E(\widehat{\mu_i} - \mu_i)^2$$

$$= E(\widetilde{\mu_i} - \mu_i)^2 + E(\widehat{\mu_i} - \widetilde{\mu_i})^2 \ , \quad i = 1, 2, \cdots, m$$

단, $\widetilde{\mu_i}$ 는 μ_i 의 BLUP 추정량이며, 여기에서 첫 번째 항인 $E(\widetilde{\mu_i} - \mu_i)^2$은 다음 (5.8)식의 결과에서 계산된다.

$$MSE(\widetilde{\mu_i}) = E(\widetilde{\mu_i} - \mu_i)^2$$

$$= \overline{X_i}^t (G_i^{-1})^t \Omega \overline{X_i}$$

$$+ \sigma^2 \overline{X_i} (G_i^{-1})^t Z_i \left(\sum_{i=1}^{m} Z_i^t G_i^{-1} X_i^t X_i Z_i \right)^{-1} Z_i^t G_i^{-1} \overline{X_i}$$

$$= T_1 + T_2 \ , \qquad (5.8)$$

단, $G_i = I + \sigma^2 X_i^t X_i \Omega$

두 번째 항 $E(\widehat{\mu_i} - \widetilde{\mu_i})^2$은 (5.9)식과 같이 근사적인 계산식으로 주어진다.

$$E(\widehat{\mu_i} - \widetilde{\mu_i})^2 \approx T_3$$

$$= tr\left[\left(\frac{\partial d_i}{\partial \theta} \right) V \left(\frac{\partial d_i}{\partial \theta} \right)^t E\{ (\hat{\theta} - \theta)(\hat{\theta} - \theta)^t \} \right] \qquad (5.9)$$

여기에서 $d_i = \overline{X_i}^t K_i (I \otimes \Omega) X^t V^{-1}$, $K_i = [0, \cdots, 0, I, 0, \cdots, 0]$는 $(p+1) \times (p+1)m$ 행렬로써 i 번째 소지역에서 $(p+1) \times (p+1)$ 단위행렬을 가지며, $\hat{\theta} = (\widehat{\theta_1}, \widehat{\theta_2}, \cdots, \widehat{\theta_s})$은 $\theta = (\theta_1, \theta_2, \cdots, \theta_s)$의 translation invariant 추정량으로써, $\theta_s = \sigma^2$, $\theta_k (k = 1, 2, \cdots, s-1)$는 Ω의 서로다른 원소를 나타낸다.

위의 (5.9)식에서 $E\{ (\hat{\theta} - \theta)(\hat{\theta} - \theta)^t \}$ 를 REML 추정량 B의 점근적 공분산

행렬로 근사시킨 계산 결과식은 다음 (5.10)식과 같다.

$$T_3 = \overline{X_i}^t (G_i^{-1}) \left(\sum_{j=1}^{s-1} \sum_{k=1}^{s-1} b_{jk} \triangle_j C_i \triangle_k{}^t \right) G_i^{-1} \overline{X_i}$$

$$- 2\, \overline{X_i}^t (G_i^{-1})^t \left(\sum_{j=1}^{s-1} b_{j,s} \triangle_j \right) R_i \Omega \overline{X_i} + b_{ss}\, \overline{X_i}^t \Omega S_i \Omega\, \overline{X_i}\,, \qquad (5.10)$$

여기에서 $B = \theta$ 의 REML 추정량, $b_{j,k} = B$ 의 jk 번째 원소, $b^*_{jk} = B^{-1}$ 의 jk 번째 원소 $= tr\left(\sum_{i=1}^{m} P_i \frac{\partial V}{\partial \theta_j} P_i \frac{\partial V}{\partial \theta_k} \right)$ $(k = 1, 2, \cdots, s)$ 이며, P_i는 다음과 같이 표현된다.

$$P_i = V_i^{-1} - V_i^{-1} X_i Z_i \left(\sum_{i=1}^{m} Z_i^t X_i^t V_i^{-1} X_i Z_i \right) Z_i^t X_i^t V_i^{-1},$$

단, $C_i = \sigma^{-2} G_i^{-1} X_i^t X_i$,

$R_i = \sigma^{-4} G_i^{-2} X_i^t X_i$,

$S_i = \sigma^{-6} G_i^{-3} X_i^t X_i$,

$\triangle_k = \frac{\partial \Omega}{\partial \Omega_k}$: $(s-1) \times (s-1)$ $(k = 1, 2, \cdots, s-1)$.

위의 (5.8)식과 (5.10)식을 정리하면, $\widehat{\mu_i}$ 의 MSE 근사식은 다음 (5.11)식과 같이 표현된다.

$$MSE(\widehat{\mu_i}) = E(\widehat{\mu_i} - \mu_i)^2$$

$$= E(\widetilde{\mu_i} - \mu_i)^2 + E(\widehat{\mu_i} - \widetilde{\mu_i})^2\,, \quad i = 1\,,\, 2\,,\, \cdots\,,\, m$$

$$\approx \left\{ \overline{X_i}^t (G_i^{-1})^t \Omega\, \overline{X_i} \right\}$$

$$+\left\{ \sigma^2 \overline{X_i}(G_i^{-1})^t Z_i \left(\sum_{i=1}^{m} Z_i^t G_i^{-1} X_i^t X_i Z_i \right)^{-1} Z_i^t G_i^{-1} \overline{X_i} \right\}$$

$$+\left\{ \overline{X_i}^t (G_i^{-1}) \left(\sum_{j=1}^{s-1} \sum_{k=1}^{s-1} b_{jk} \triangle_j C_i \triangle_k^t \right) G_i^{-1} \overline{X_i} \right.$$

$$\left. -2\, \overline{X_i}^t (G_i^{-1})^t \left(\sum_{j=1}^{s-1} b_{j,s} \triangle_j \right) R_i \Omega \overline{X_i} + b_{ss}\, \overline{X_i}^t \Omega S_i \Omega \overline{X_i} \right\}$$

$$= T_1 + T_2 + T_3 \qquad (5.11)$$

한편, 위의 결과들을 유한 모집단에 적용할 경우에는 다음과 같이 보정 절차를 거쳐 표본으로 추출되지 않은 단위들의 영향력을 고려해 주어야 한다.

$$\hat{\mu}_i^F = f_i\, \bar{y}_i + (\overline{X_i} - f_i\, \bar{x}_i)^t (Z_i\, \hat{\gamma} + \hat{\nu}_i)\ , \qquad (5.12)$$

단, $f_i = n_i / N_i$, $\bar{x}_i = (1\ ,\ \bar{x}_{i1}\ ,\ \bar{x}_{i2}\ ,\ \cdots\ ,\ \bar{x}_{ip})^t$ 는 표본평균을 나타낸다.

$$\hat{\mu}_i^F - \overline{Y} = (1 - f_i)\left[(\ \overline{X}_i^C)^t \left\{ Z_i(\hat{\gamma} - \gamma) + \hat{\nu}_i - \nu_i - \ \bar{\varepsilon}_i^C \right\} \right]\ , \qquad (5.13)$$

여기에서 $\overline{X}_i^C = (1 - f_i)^{-1} (\overline{X_i} - f_i\, \bar{x}_i)$, $\bar{\varepsilon}_i^C$ = 'i 번째 소지역에서 표본으로 추출되지 않은 단위들에 대한 ε_{ij} 의 평균'을 나타낸다.

위의 (5.12)식과 (5.13)식으로부터 추정량의 MSE를 계산하면 다음 (5.14)식과 같이 주어진다.

$$MSE(\ \hat{\mu}_i^F) = (1 - f_i)^2 \left\{ MSE^*(\hat{\mu}_i) + N_i^{-1}(1 - f_i)^{-1} \sigma^2 \right\}\ , \qquad (5.14)$$

여기에서 $MSE^*(\hat{\mu}_i)$는 (5.11)식에서 $\overline{X_i}$ 를 $\overline{X}_i^C$ 로 대체하여 구한다.

$\hat{\mu}_i$ 의 MSE 추정식은 $\hat{\mu}_i$ 의 MSE 근사식 (5.11)과 (5.14))식에서 σ^2과 Ω 의 RIGLS 추정량을 이용할 수 있으며, RIGLS 추정량을 이용한 $\hat{\mu}_i$ 의 MSE 추정식은 (5.15)와 (5.16)과 같이 주어진다.

$$\widehat{MSE}(\hat{\mu}_i) = \hat{T}_1 + \hat{T}_2 + \hat{T}_3 , \tag{5.15}$$

$$\widehat{MSE}(\hat{\mu}_i^F) = (1-f_i)^2 \{ \widehat{MSE}^*(\hat{\mu}_i) + N_i^{-1}(1-f_i)^{-1} \hat{\sigma}^2 \}. \tag{5.16}$$

5.1.4 소지역 추정에 적용

Moura et al.(1999)는 Multilevel 모형이 소지역 추정에 광범위하게 적용될 수 있는 가능성을 최근 논문에 발표하였다. 브라질의 하나의 county내의 전체 지역에서 조사된 자료를 이용하여 Multilevel 모형하에서 미지인 모수들을 앞서 소개하였던 방법에 의해 추정하고, 전체 자료와 전체 자료 중 일부분을 랜덤 추출하여 얻은 표본 자료를 이용하여 소지역의 평균에 관한 추정치의 정확도를 비교하여 소지역 추정에 Multilevel 모형이 적용될 수 있다는 가능성을 확인해 주었다.

사용된 자료는 브라질의 한 County내의 전체 지역에서 조사된 38,740 가구에 대한 조사자료이며, 종속변수는 가장의 수입액, 설명변수는 가장의 교육정도(0 ~ 5)와 가구별 방의 수(1 ~ 11)이다.

(1) 모수 추정

소지역 특성치를 추정하기 위해 사용된 Multilevel 모형은 다음과 같다.

$$Y_{ij} = \beta_{i0} + \beta_{i1} x_{1ij} + \beta_{i2} x_{2ij} + \varepsilon_{ij} , \tag{5.17}$$

$$\beta_{i0} = \gamma_{00} + \nu_{i0} , \ \beta_{i1} = \gamma_{10} + \nu_{i1} , \ \beta_{i2} = \gamma_{20} + \nu_{i2} ,$$

단, $i = 1, 2, \cdots, m$(소지역 수), $j = 1, 2, \cdots, N_i$ (i 번째 소지역의 전체 조사 가구수), $\varepsilon_{ij} \overset{ind}{\sim} N(0, \sigma^2)$, $\nu_i = (\nu_{i0}, \nu_{i1}, \nu_{i2}) \overset{ind}{\sim} N(0, \Omega)$, x_{1ij} 는

방의 수, x_{2ij} 는 가장의 교육정도를 나타내며 오차항 ε_{ij}와 ν_i는 서로 독립임을 가정한다. 여기에서 x_{1ij} 와 x_{2ij} 는 각각의 모평균에 대해서 표준화된 값을 나타낸다.

county내의 전체자료를 이용하여 앞서 소개되었던 방법에 근거하여 고정 모수 γ 와 분산성분 모수 $\theta = ([\ \mathrm{Vech}(\Omega)]^t , \sigma^2)^t$ 의 RIGLS 추정량을 계산하면 결과는 다음과 같다.

i) 고정 모수 γ

$\gamma_{00} = 8.456$ (표준오차 : 0.108), $\gamma_{10} = 1.223$ (0.046) ,

$\gamma_{20} = 2.596$ (0.086)

ii) 분산성분 모수 θ

$\sigma^2 = 47.74$ (0.345) ,

$$V(\nu_i) = \begin{pmatrix} \sigma_{00} & \sigma_{01} & \sigma_{02} \\ \sigma_{10} & \sigma_{11} & \sigma_{12} \\ \sigma_{20} & \sigma_{21} & \sigma_{22} \end{pmatrix}$$

$$= \begin{pmatrix} 1.385(0.194) & 0.345(0.66) & 0.492(0.117) \\ & 0.234(0.35) & 0.333(0.054) \\ & & 0.926(0.124) \end{pmatrix}$$

(2) 모의실험을 통한 모형 적용 검토

수치조사를 시행하기 위하여 먼저 다음과 같은 자료생성 모형을 정의하기로 한다.

i) 일반 모형(G)

(5.17) 모형에서 $\nu_i = (\nu_{i0}, \nu_{i1}, \nu_{i2}) \overset{ind}{\sim} N(0, \Omega)$ 이고,

$$\Omega = \begin{pmatrix} 1.385(0.194) & 0.345(0.66) & 0.492(0.117) \\ & 0.234(0.35) & 0.333(0.054) \\ & & 0.926(0.124) \end{pmatrix}$$ 인 모형

ii) 대각 모형(D)

(5.17) 모형에서 $\nu_i = (\nu_{i0}, \nu_{i1}, \nu_{i2}) \overset{ind}{\sim} N(0, \Omega)$ 이고,

$$\Omega = \begin{pmatrix} 1.385(0.194) & 0 & 0 \\ 0 & 0.234(0.35) & 0 \\ 0 & 0 & 0.926(0.124) \end{pmatrix}$$ 인 모형

iii) 랜덤 절편항 모형(RI)

(5.17) 모형에서 $\nu_i = (\nu_{i0}, \nu_{i1}, \nu_{i2}) \overset{ind}{\sim} N(0, \Omega)$ 이고,

$$\Omega = \begin{pmatrix} 1.385(0.194) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$ 인 모형

가정된 자료생성 모형 m_1(= G, D, RI 모형)의 각각에 대해서 각 소지역별 평균의 기대값을 다음 식을 이용하여 반복 생성하고,

$$\mu_{i m_1}^{(r)} = \beta_{0i}^{(r)} + \beta_{1i}^{(r)} \overline{X}_{1i} + \beta_{2i}^{(r)} \overline{X}_{2i},$$

단, $r = 1, 2, \cdots, R(=5000)$,

$\overline{X}_{1i}$, $\overline{X}_{2i}$: 각 소지역 전체 조사자료에 대한 평균

비교를 위하여 각 소지역의 전체 조사자료에 대해서 10%의 Simulation Subset을 추출하여 위에서 정의된 각 모형별로 Y_{ij} 를 생성하고, 모형을 적합시켜 각 소지역별 평균의 기대값에 대한 추정치를 다음 식을 이용하여 반복 생성하여,

$$\hat{\mu}_{im_1}^{(r)} = \hat{\beta}_{0i}^{(r)} + \hat{\beta}_{1i}^{(r)} \bar{x}_{1i} + \hat{\beta}_{2i}^{(r)} \bar{x}_{2i} ,$$

단, $r = 1, 2, \cdots, R(=5000)$,

$\bar{x}_{1i}$, $\bar{x}_{2i}$: 각 소지역의 Simulation Subset자료에 대한 평균

$\hat{\mu}_{im_1}^{(r)}$ 과 $\mu_{im_1}^{(r)}$ 의 차에 대한 양을 비교하고자 한다.

자료생성 모형 m_1 (G, D, RI 모형)에 의해 생성된 자료를 이용하여 적합시킨 추정모형을 m_2 (G, D, RI 추정모형)라 했을 때, 다음의 측도를 정의하자.

$$MSE(\hat{\mu}_{im_1, m_2}) = \frac{\sum_{r=1}^{R} (\hat{\mu}_{im_1, m_2}^{(r)} - \mu_{im_1}^{(r)})^2}{R} ,$$

$$ARE(\hat{\mu}_{im_1, m_2}) = \frac{\sum_{r=1}^{R} |\hat{\mu}_{im_1, m_2}^{(r)} - \mu_{im_1}^{(r)}| / \mu_{im_1}^{(r)}}{R} ,$$

$$\text{RMSE}_{m_2, m_1} = \sum_{i=1}^{m} \frac{MSE(\hat{\mu}_{im_1, m_2})}{MSE(\hat{\mu}_{im_1, m_1})} \times 100 ,$$

$$RARE_{m_2, m_1} = \sum_{i=1}^{m} \frac{ARE(\hat{\mu}_{im_1, m_2})}{ARE(\hat{\mu}_{im_1, m_1})} \times 100 .$$

여기에서 $MSE(\hat{\mu}_{im_1, m_2})$와 $ARE(\hat{\mu}_{im_1, m_2})$는 전체자료 중 10%의 Simulation Subset자료를 이용하여 m_1 모형에서 자료를 생성하여 m_2 모형을 적합시켰을 때

i소지역의 평균에 관한 추정치의 평균제곱오차 및 절대상대오차를 각각 의미하며, RMSE_{m_2,m_1}과 $RARE_{m_2,m_1}$는 m_1 모형에서 자료를 생성하여 m_1 모형을 추정했을때와 m_1 모형에서 자료를 생성하여 m_2 모형을 추정했을때의 추정량의 값에 대한 평균제곱오차비, 절대허용오차비를 각각 의미한다.

모의실험의 절차를 단계적으로 살펴보면 다음과 같이 설명할 수 있다.

제 1단계

Step1

일반모형(G)에서 명시된 $\nu_i = (\nu_{i0}, \nu_{i1}, \nu_{i2})$ 의 분포로부터 $\nu_1, \nu_2, \cdots, \nu_m$을 생성한다.

Step2

Step1에서 생성된 ν_i ($i = 1, 2, \cdots, m$) 에 대해서 $\beta_1, \beta_2, \cdots, \beta_m$ 을 결정한다. 여기에서 $\beta_i = (\beta_{i0}, \beta_{i1}, \beta_{i2})$, $\beta_{i0} = \gamma_{00} + \nu_{i0}$, $\beta_{i1} = \gamma_{10} + \nu_{i1}$, $\beta_{i2} = \gamma_{20} + \nu_{i2}$ 이며, 고정모수 $\gamma_{00} = 8.456$, $\gamma_{10} = 1.223$, $\gamma_{20} = 2.596$는 이미 주어진 값이다.

Step3

이미 계산된 i 번째 소지역의 전체 조사자료에 대한 평균들 ($\overline{X}_{1i}$, $\overline{X}_{2i}$)와 Step2에서 결정된 $\beta_i = (\beta_{i0}, \beta_{i1}, \beta_{i2})$를 사용하여 일반모형(G)에 대한 i 번째 소지역의 평균에 대한 기대값 $\mu_{iG}{}^{(1)}$을 생성한다.

$$\mu_{iG}{}^{(1)} = \beta_{0i} + \beta_{1i}\,\overline{X}_{1i} + \beta_{2i}\,\overline{X}_{2i}\,,$$

여기에서 $i = 1, 2, \cdots, m$(소지역 수) , G 는 일반모형을 나타낸다. 즉 일반모형에

대해서 $\mu_{1G}^{(1)}$, $\mu_{2G}^{(1)}$, $\cdots$, $\mu_{mG}^{(1)}$이 생성된다.

Step4

Step1 ~ Step3 의 과정을 5000번 반복하여 General Model 의 i 번째 소지역의 평균에 대한 기대값 μ_{iG} 를 다음 식을 이용하여 5000개씩 생성한다.

$$\mu_{iG}^{(r)} = \beta_{0i}^{(r)} + \beta_{1i}^{(r)} \overline{X}_{1i} + \beta_{2i}^{(r)} \overline{X}_{2i},$$

단, $i = 1, 2, \cdots, m$ (소지역 수), $r = 1, 2, \cdots, R(=5000)$

Step5

대각모형(D), 랜덤절편항 모형(RI)에 대해서도 위의 Step1~Step4를 5000번 반복하여 $\mu_{iD}^{(r)}$, $\mu_{iRI}^{(r)}$을 생성한다.

제 2단계

Step1

제 1단계의 Step1 ~ Step2에서 얻어진 $\beta_1, \beta_2, \cdots, \beta_m$ 을 취한다.

Step2

각각의 소지역으로부터 10%의 표본을 랜덤 추출(=Simulation Subset)하여 보조정보 (x_{1ij}, x_{2ij})를 확보하고, $\varepsilon_{ij} \sim N(0, \sigma^2 = 47.74)$로부터 보조 정보의 개수만큼 ε_{ij} 를 생성한다. 여기에서 $j(=1, 2, \cdots, n_i)$는 i 번째 소지역에서 추출된 Simulation Subset의 개수를 나타낸다.

Step3

Step1~Step2에서 생성된 β_i, (x_{1ij}, x_{2ij}), ε_{ij} 값을 다음 식에 대입하여 n_i

개의 Y_{ij} 값을 생성한다.

$$Y_{ij} = \beta_{i0} + \beta_{i1} x_{1ij} + \beta_{i2} x_{2ij} + \varepsilon_{ij}$$

Step4

i 번째 소지역에서 랜덤 추출된 n_i 개의 (x_{1ij} , x_{2ij}) 값과 이에 대응되는 n_i 개의 Y_{ij} 값을 이용하여 다음과 같은 세가지 모형(일반모형(G), 대각모형(D), 랜덤절편항 모형(RI))을 추정함

o 일반모형(G)에서 자료를 생성하여 일반모형(G)을 추정

$$\hat{Y}_{iG,G}^{(1)} = \hat{\beta}_{i0}^{(1)} + \hat{\beta}_{i1}^{(1)} x_{1i} + \hat{\beta}_{i2}^{(1)} x_{2i}$$

o 일반모형(G)에서 자료를 생성하여 대각모형(D)을 추정

$$\hat{Y}_{iG,D}^{(1)} = \hat{\beta}^{*}_{i0}{}^{(1)} \hat{\beta}^{*}_{i1}{}^{(1)} x_{1i} + \hat{\beta}^{*}_{i2}{}^{(1)} x_{2i}$$

o 일반모형(G)에서 자료를 생성하여 랜덤절편항 모형(RI)을 추정

$$\hat{Y}_{iG,RI}^{(1)} = \hat{\beta}^{**}_{i0}{}^{(1)} \hat{\beta}^{**}_{i1}{}^{(1)} x_{1i} + \hat{\beta}^{**}_{i2}{}^{(1)} x_{2i}$$

Step5

Step4 에서 계산된 회귀계수의 추정치를 이용하여 세가지 모형에 대해서 i 번째 소지역의 평균에 대한 기대값의 추정치를 다음 식으로 계산한다.

o 일반모형(G)에서 자료를 생성하여 일반모형(G)을 추정한 경우

$$\hat{\mu}_{iG,G}^{(1)} = \hat{\beta}_{i0}^{(1)} + \hat{\beta}_{i1}^{(1)} \bar{x}_{1i} + \hat{\beta}_{i2}^{(1)} \bar{x}_{2i}$$

o 일반모형(G)에서 자료를 생성하여 대각모형(D)을 추정한 경우

$$\hat{\mu}_{iG,D}^{(1)} = \hat{\beta}^{*}_{i0}{}^{(1)} \hat{\beta}^{*}_{i1}{}^{(1)} \bar{x}_{1i} + \hat{\beta}^{*}_{i2}{}^{(1)} \bar{x}_{2i}$$

o 일반모형(G)에서 자료를 생성하여 랜덤절편항 모형(RI)을 추정한 경우

$$\hat{\mu}_{iG,RI}^{(1)} = \hat{\beta}^{**}_{i0}{}^{(1)} \quad \hat{\beta}^{**}_{i1}{}^{(1)} \bar{x}_{1i} + \hat{\beta}^{**}_{i2}{}^{(1)} \bar{x}_{2i}$$

여기에서 $\bar{x}_{1i}$; $\bar{x}_{2i}$ 는 Simulation Subset으로 추출된 x_{1ij} , x_{2ij} 의 평균을 나타낸다.

Step6

Step1 ~ Step5 의 절차를 5000번 반복하여 $\hat{\mu}_{iG,G}^{(r)}$, $\hat{\mu}_{iG,D}^{(r)}$, $\hat{\mu}_{iG,RI}^{(r)}$를 생성한다.

Step7

o 대각모형(D)을 가정하여 Step1의 절차에 따라 β_1 , β_2 , … , β_m 을 취하고, Step2 ~Step6의 절차를 수행하여 $\hat{\mu}_{iD,G}^{(r)}$, $\hat{\mu}_{iD,D}^{(r)}$, $\hat{\mu}_{iD,RI}^{(r)}$를 생성한다.

o 랜덤절편항 모형(RI)을 가정하여 Step1의 절차에 따라 β_1 , β_2 , … , β_m 을 취하고, Step2~Step6의 절차를 수행하여 $\hat{\mu}_{iRI,G}^{(r)}$, $\hat{\mu}_{iRI,D}^{(r)}$, $\hat{\mu}_{iRI,RI}^{(r)}$ 를 생성한다.

제 3단계

o $MSE(\hat{\mu}_{im_1,m_2}) = \dfrac{\sum_{r=1}^{R} (\hat{\mu}_{im_1,m_2}^{(r)} - \mu_{im_1}^{(r)})^2}{R}$ 계산,

; i.e $MSE(\hat{\mu}_{iG,G})$, $MSE(\hat{\mu}_{iG,D})$, $MSE(\hat{\mu}_{iG,RI})$,

$MSE(\hat{\mu}_{iD,G})$, $MSE(\hat{\mu}_{iD,D})$, $MSE(\hat{\mu}_{iD,RI})$,

$MSE(\hat{\mu}_{iRI,G})$, $MSE(\hat{\mu}_{iRI,D})$, $MSE(\hat{\mu}_{iRI,RI})$ 를 계산

$ARE(\hat{\mu}_{im_1,m_2}) = \dfrac{\sum_{r=1}^{R} |\hat{\mu}_{im_1,m_2}{}^{(r)} - \mu_{im_1}{}^{(r)}| / \mu_{im_1}{}^{(r)}}{R}$ 계산

; i.e $ARE(\hat{\mu}_{iG,G})$, $ARE(\hat{\mu}_{iG,D})$, $ARE(\hat{\mu}_{iG,RI})$,

$ARE(\hat{\mu}_{iD,G})$, $ARE(\hat{\mu}_{iD,D})$, $ARE(\hat{\mu}_{iD,RI})$,

$ARE(\hat{\mu}_{iRI,G})$, $ARE(\hat{\mu}_{iRI,D})$, $ARE(\hat{\mu}_{iRI,RI})$ 를 계산

o $\mathrm{RMSE}_{m_2,m_1} = \sum_{i=1}^{m} \dfrac{MSE(\hat{\mu}_{im_1,m_2})}{MSE(\hat{\mu}_{im_1,m_1})} \times 100$ 을 계산,

; i.e $\mathrm{RMSE}_{G,G}$, $\mathrm{RMSE}_{G,D}$, $\mathrm{RMSE}_{G,RI}$,

$\mathrm{RMSE}_{D,G}$, $\mathrm{RMSE}_{D,D}$, $\mathrm{RMSE}_{D,RI}$,

$\mathrm{RMSE}_{RI,G}$, $\mathrm{RMSE}_{RI,D}$, $\mathrm{RMSE}_{RI,RI}$ 를 계산

$RARE_{m_2,m_1} = \sum_{i=1}^{m} \dfrac{ARE(\hat{\mu}_{im_1,m_2})}{ARE(\hat{\mu}_{im_1,m_1})} \times 100$ 을 계산

; i.e $RARE_{G,G}$, $RARE_{G,D}$, $RARE_{G,RI}$,

$RARE_{D,G}$, $RARE_{D,D}$, $RARE_{D,RI}$,

$RARE_{RI,G}$, $RARE_{RI,D}$, $RARE_{RI,RI}$ 를 계산

모의실험 결과를 요약하면 다음의 <표1>과 같다.

<표 1> 자료생성 모형 $m_1 = G, D, RI$ 와 추정모형 $m_2 = G, D, RI$ 에 대한 RMSE ($RARE$) 값의 비교

Estimator	Data Generation Model		
	일반모형(G)	대각모형(D)	랜덤절편항모형 (RI)
General (G)	100.0 * (100.0)	101.8 ** (100.9)	101.2 (100.6)
Diagonal(D)	108.8 (82.6)	100.0 (100.0)	100.2 (100.1)
Random Intercept (RI)	131.9 (176.9)	109.1 (105.6)	100.0 (100.0)

$$(*)\quad \mathrm{RMSE}_{G,G} = \sum_{i=1}^{m} \frac{MSE(\hat{\mu}_{iG,G})}{MSE(\hat{\mu}_{iG,G})} \times 100 = 100.0$$

$$(**)\quad \mathrm{RMSE}_{G,D} = \sum_{i=1}^{m} \frac{MSE(\hat{\mu}_{iD,G})}{MSE(\hat{\mu}_{iD,D})} \times 100$$

$$= \frac{\sum_{i=1}^{m}\left\{\frac{\sum_{r=1}^{5000}(\hat{\mu}_{iD,G}{}^{(r)} - \mu_{iD}{}^{(r)})^2}{R}\right\}}{\sum_{i=1}^{m}\left\{\frac{\sum_{r=1}^{5000}(\hat{\mu}_{iD,D}{}^{(r)} - \mu_{iD}{}^{(r)})^2}{R}\right\}} \times 100 = 101.8$$

랜덤절편항 모형(RI)과 같은 단순한 모형에서 자료를 생성하여, 이 모형보다 복잡한 일반모형(G), 대각모형(D)을 적합시켜 얻은 추정량들의 평균제곱오차비 (RMSE)를 살펴보면, $\mathrm{RMSE}_{G,RI} = 101.2$, $\mathrm{RMSE}_{D,RI} = 100.2$로써 심한 효율의 손실은 발생하지 않았음을 알 수 있다. 이와는 반대로 일반모형(G)과 같은 비교적 복잡한 모형에서 자료를 생성하여, 이 모형보다는 단순한 대각모형(D), 랜덤절편항 모형(RI) 적합시켜 얻은 RMSE값을 살펴보면, $\mathrm{RMSE}_{D,G} = 108.8$, $\mathrm{RMSE}_{RI,G} = 131.9$ 로써 추정의 효율이 떨어짐을 확인할 수 있다.

RMSE 값 간의 차이를 살펴보면, RMSE_{G,m_1} 와 RMSE_{RI,m_1} 간의 차이는 크

게 나타난다. 특히 $RMSE_{G,G} = 100.0$와 $RMSE_{RI,G} = 131.9$ 간의 차이와 $RMSE_{G,D} = 101.8$와 $RMSE_{RI,D} = 109.1$ 간의 차이가 크게 나타난다. 또한 $RMSE_{D,m_1}$과 $RMSE_{RI,m_1}$간의 차이도 크게 나타나며, 세부적으로 살펴보면 $RMSE_{D,G} = 108.8$와 $RMSE_{RI,G} = 131.9$ 간의 차이와 $RMSE_{D,D} = 100.0$와 $RMSE_{RI,D} = 109.1$ 간의 차이가 크다. 반면 $RMSE_{G,m_1}$와 $RMSE_{D,m_1}$간의 차이는 전체적으로는 작게 나타난다.

이상의 결과를 요약하면 주어진 자료에 대해서 소지역 추정시 랜덤절편항 모형(RI)을 가정하는 것은 될 수 있다면 피하는 것이 바람직하며, 그러나 반드시 복잡한 일반모형(G)만을 고집할 필요는 없겠고, 일반모형(G)이나 대각모형(D) 중 어떤 것을 선택하여도 무난할 것으로 판단된다. 더불어 소지역 추정의 효율을 높일 수 있는 추가적인 랜덤계수의 도입도 고려해 볼만 연구 사항이다.

회귀모형에 비해서 Multilevel 모형은 소지역들이 지역들 간의 특성을 보유하면서 하나의 통일된 형태로 모형이 표현된다는 것을 장점으로 들 수 있다. 각각의 소지역에서 표본의 수가 적을 때 Multilevel 모형을 이용하여 추정하였을 경우 좋은 결과를 예측하기가 어려울 것으로 생각할 수 있으나, 이러한 경우에도 Multilevel 모형을 이용한 추정이 소지역별로 분리되어 추정된 회귀모형 추정보다 평균적으로 더 좋은 결과를 보여 준다. 다음의 <표 2>는 이러한 결과를 설명한다.

<표 2> RMSE 와 RARE

Estimator	Data Generation Model	
	General Model	Separate Regression Model
G	100.0 (100.0)	88.1 (83.1)
Separate Regression	247.6 (154.7)	100.0 (100.0)

Multilevel 모형의 자료 생성 시 앞에서 계산된 모수값을 이용하였고, 일반적 회귀모형에서 자료 생성시 모수의 추정값 $\hat{\sigma}^2$ 은 각각의 소지역에서 추정된 값을 이용하였다. 소지역 간의 랜덤효과를 고려하지 않은 일반적 회귀모형을 이용했을 경우 소지역 추정은 심각한 효율의 손실을 가져올 수 있음을 확인 할 수 있다. 한편, $\mathrm{RMSE}_{SR,G}$ 값을 각각의 소지역별로 분리하여 살펴보면, 소지역의 표본크기가 클수록 작은 경향을 보인다. 소지역의 표본크기가 큰 경우에는 $\mathrm{RMSE}_{G,G}$ 값과 $\mathrm{RMSE}_{SR,G}$ 값 간의 차이는 줄어드는 경향을 보인다.

다음은 모의실험의 결과로부터 $\hat{\mu}_i$ 의 근사 MSE 성질을 설명하고자 한다.

비교기준으로 $\hat{\mu}_i$ 의 MSE 근사식은 다음 식을 이용하고,

$$
\begin{aligned}
MSE(\hat{\mu}_i) &= E(\hat{\mu}_i - \mu_i)^2 \\
&= E(\tilde{\mu}_i - \mu_i)^2 + E(\hat{\mu}_i - \tilde{\mu}_i)^2 \ , i=1,2,\cdots,m \\
&\approx \left\{ \overline{X}_i^t (G_i^{-1})^t \Omega \overline{X}_i \right\} \\
&+ \left\{ \sigma^2 \overline{X}_i (G_i^{-1})^t Z_i \left(\sum_{i=1}^{m} Z_i^t G_i^{-1} X_i^t X_i Z_i \right)^{-1} Z_i^t G_i^{-1} \overline{X}_i \right\} \\
&+ \left\{ \overline{X}_i^t (G_i^{-1}) \left(\sum_{j=1}^{s-1} \sum_{k=1}^{s-1} b_{jk} \wedge_j C_i \triangle_k^t \right) G_i^{-1} \overline{X}_i \right. \\
&\left. - 2 \overline{X}_i^t (G_i^{-1})^t \left(\sum_{j=1}^{s-1} b_{j,s} \triangle_j \right) R_i \Omega \overline{X}_i + b_{ss} \overline{X}_i^t \Omega S_i \Omega \overline{X}_i \right\} , \\
&= T_1 + T_2 + T_3
\end{aligned}
$$

모의실험에 의한 $\hat{\mu}_i$ 의 MSE 근사값으로는 다음의 식

$$MSE(\hat{\mu}_{iG}^{(r)}) = E(\hat{\mu}_{iG,G}^{(r)} - \mu_{iG}^{(r)})^2$$

을 이용한다. 단, $r=1,2,\cdots,R\,(=5000)$ 이고, T_1 , T_2 , T_3 의 X_i 값은

10%의 Simulation Subset 의 값으로 대체된다.

일반모형(G)에서 자료를 생성하여 일반모형(G)을 추정한 경우 MSE 의 근사는 매우 좋게 나타났다. 5000번의 모의실험 결과로부터 생성된 $MSE(\hat{\mu}_{iG}^{(r)})$의 값을 $MSE(\hat{\mu}_i)$ 값과 비교해 보면, 평균적인 과소추정의 양은 $MSE(\hat{\mu}_i)$값의 0.31%정도이고, 이러한 과소추정의 양 중 가장 큰 값은 $MSE(\hat{\mu}_i)$값의 약 5.4%정도이며, 과대추정의 양 중 가장 큰 값은 $MSE(\hat{\mu}_i)$값의 약 4.8%정도로 나타났다.

모의실험 결과 T_1은 평균적으로 $MSE(\hat{\mu}_i)$의 94.6% 정도, T_3는 평균적으로 $MSE(\hat{\mu}_i)$의 4.3%정도를 차지한다. 각각의 소지역에서 살펴보면 T_1은 $MSE(\hat{\mu}_i)$의 87.4%에서 99.1%의 범위에 있고, T_3는 0.7%에서 10.5%의 범위에 있으며, T_2는 $MSE(\hat{\mu}_i)$값의 2.2%미만의 범위에 있음이 확인되었다.

마지막으로 $MSE(\hat{\mu}_i)$의 추정결과를 살펴보기로 한다. 5000번의 모의실험 결과로부터 계산된 $\widehat{MSE}(\hat{\mu}_{iG}^{(r)}) = \hat{T}_1^* + \hat{T}_2^* + 2\hat{T}_3^*$의 값을 $\widehat{MSE}(\hat{\mu}_i) = \hat{T}_1 + \hat{T}_2 + 2\hat{T}_3$의 값과 비교해 보면 근사적으로 Unbiased 되어 있고, $\widehat{MSE}(\hat{\mu}_{iG}^{(r)}) = \hat{T}_1^* + \hat{T}_2^*$값을 $\widehat{MSE}(\hat{\mu}_i) = \hat{T}_1 + \hat{T}_2$과 비교해 보면, 평균적으로 약 9.1%정도 과소추정됨이 확인된다. 이는 Singh et al.(1988), Prasad and Rao(1990)의 결과와도 일치한다.

소지역 추정시 특정 분산성분을 갖는 모형이 소지역 추정의 정확도를 개선할 수 있다는 사실을 제안한 연구는 Prasad and Rao(1990), Battese et. al(1981, 1988)에 의해서였다. Moura et al.(1999)는 특정 분산항을 갖는 모형인 Multilevel 모형이 소지역 추정의 정확도를 개선할 수 있다는 적용사례를 제시하여 소지역 추정에 관한 연구를 가일층 진보시켰다.

Moura et al.(1999)의 모의실험 결과로부터 확인할 수 있는 사항은 다음과 같이 요약된다. 첫째, 랜덤절편항 모형(RI)보다는 좀더 복잡한 구조를 갖는 대각모형(D), 일반모형(G)과 같은 분산성분 모형을 이용할 경우 소지역 추정의 정확도가 개선될 수 있다. 둘째, 분산성분을 갖는 모형을 가정할 때, 좀 더 복잡한 모형을 가정하더라도 추정의 효율은 심하게 떨어지지 않는다. 셋째, 소지역 공변량을 도입하여 소지역 추정의 정확도를 개선할 수 있다. 넷째, 소지역 추정시 Multilevel 모형 이용이 일상적인 회귀모형 이용보다 오히려 선호 되어야 한다.

한편, Moura et al(1999)의 모의실험 결과를 살펴보면, $\widehat{\mu_i}$ 의 *MSE* 근사는 정도가 높게 나타나고 *MSE* 추정은 근사적으로 Unbiased 되어 있으나 추가적으로 근사의 정확한 order 에 대한 이론적인 연구가 계속 이루어져야 하고, 소지역 추정의 정확도를 높일 수 있는 일반혼합효과 모형의 적용 연구도 진행되어야 할 사항이다.

5.2 Unit-Level 모형을 이용한 소지역 추정 사례(II)

5.2.1 서 론

미 농림부에서는 농작물 재배면적의 예측을 보다 정확히 하기 위하여 지역관측위성(LANDSAT)을 이용한 위성자료를 이용해오고 있다. 여기에서는 1978년 6월 Iowa 12개 County에 대하여 옥수수와 콩의 재배면적에 관한 사전 조사자료와, 8월과 9월 사이에 지역위성으로부터 관측된 옥수수와 콩의 실제 재배면적자료를 근거로하여 각 County에 대하여 옥수수와 콩의 재배면적에 관한 예측 문제를 가정된 단위수준(Unit-level) 모형에 근거하여 다뤘다. 이용된 자료는 미 농림부(USDA)에서 제시한 자료로써 <표 3>에 소개하였으며, 12개 County 중 37개 구역(Segment)의 조사자료만을 분석에 이용하였다.

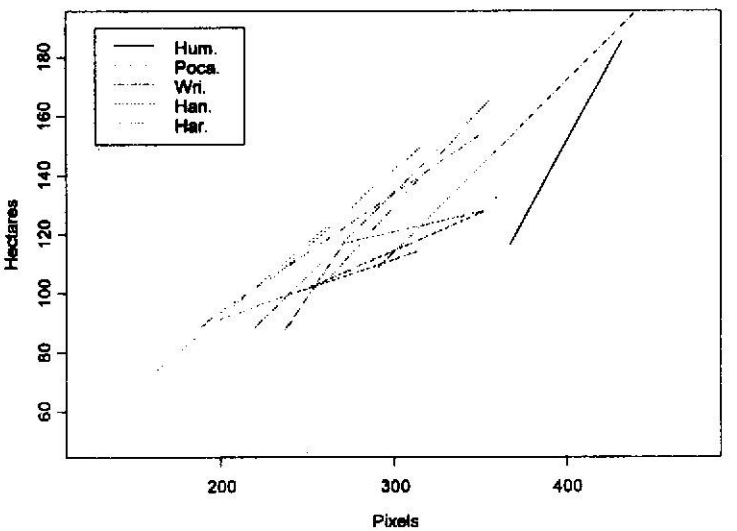
<표 3> Survey and Satellite Data for Corn and Soybeans in 12 Iowa Couties

	No. of segment		Reported hectares		No. of pixels in sample segments		Mean No. of pixels per segment	
county	Sample	County	Corn	Soybeans	Corn	Soybeans	Corn	Soybeans
Cerro Gordo	1	545	165.76	8.09	374	55	295.29	189.70
Hamilton	1	566	96.32	106.03	209	218	300.40	196.65
Worth	1	394	76.08	103.60	253	250	289.60	205.28
Humbolt	2	424	185.35	6.47	432	96	290.74	220.22
			116.43	63.82	367	178		
Franklin	3	564	162.08	43.50	361	137	318.21	188.06
			152.04	71.43	288	206		
			161.75	42.49	369	165		
Pocahontas	3	570	92.88	105.26	206	218	257.17	247.13
			149.94	76.49	316	221		
			64.75	174.34	145	338		
Winnebago	3	402	127.07	95.67	355	128	291.77	185.37
			133.55	76.57	295	147		
			77.70	93.48	223	204		
Wright	3	567	206.39	37.84	459	77	301.26	221.36
			108.33	131.12	290	217		
			118.17	124.44	307	258		
Webster	4	687	99.96	144.15	252	303	262.17	247.09
			140.43	103.60	293	221		
			98.95	88.59	206	222		
			131.04	115.58	302	274		
Hancock	5	569	114.12	99.15	313	190	314.28	198.66
			100.60	124.56	246	270		
			127.88	110.88	353	172		
			116.90	109.14	271	228		
			87.41	143.66	237	297		
Kossuth	5	965	93.48	91.05	221	167	298.65	204.61
			121.00	132.33	369	191		
			109.91	143.14	343	249		
			122.66	104.13	342	182		
			104.21	118.57	294	179		
Hardin	6	556	88.59	102.59	220	262	325.99	177.05
			88.59	***29.46***	***340***	***87***		
			165.35	69.28	355	160		
			104.00	99.15	261	221		
			88.63	143.66	187	345		
			153.70	94.49	350	190		

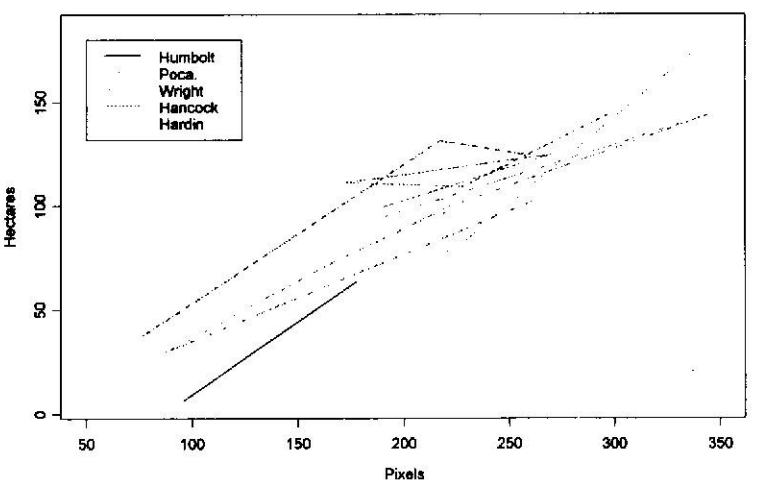
* Hardin의 자료중 하나의 관측점은 이상치로 간주하고 분석에서 제외됨

County별로 옥수수와 콩의 재배면적에 관한 조사자료와 위성에 의해 관측된 Pixel수에 관한 상관성을 살펴보기 위해 각각에 대해 산점도를 그려 직선으로 연결해 보면 <Fig1>과 <Fig2>와 같다.

<Fig 1> Corn ha vs pixels by county



<Fig 2> Soybean ha vs pixels by county



옥수수와 콩의 재배면적(ha)과 Pixel 과의 산점도에서는 강한 상관성을 보이며, 따라서 Model을 가정할 때 하나의 County내에서의 편차들은 서로 상관성이 있는 것으로 가정하는 것이 합당하며, 또한 Model의 랜덤오차들은 Nested-Error Model에 의해 정의되는 것으로 가정한다.

5.2.2 분산 성분 모형(Components-of-variance Models)

가정된 분산성분 모형은 다음의 (5.18)과 같다.

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + u_{ij} \tag{5.18}$$

단, $i = 1, 2, \cdots, T$ (county 수, $T = 12$), $j = 1, 2, \cdots, n_i$ (i번째 county에서 sample segment 수), y_{ij}= i번째 county에서 j번째 sample segment에 있는 옥수수(또는 콩)의 조사된 재배면적(ha), x_{1ij}= i번째 county에서 j번째 segment에 있는 옥수수의 pixel 수, x_{2ij}= 콩의 pixel 수, u_{ij}는 다음과 같이 가정한다.

$$u_{ij} = \nu_i + \epsilon_{ij} \tag{5.19}$$

여기에서 ν_i= i번째 county 효과 $\overset{iid}{\sim} N(0, \sigma_\nu^2)$ ($i = 1, 2, \cdots, T$), ϵ_{ij}= i번째 county에서 j번째 sample segment와 관련된 랜덤 효과 $\overset{iid}{\sim} N(0, \sigma_e^2)$를 가정하며, 이때 공분산 구조는 다음의 (5.20)과 같다.

$$E(u_{ij} u_{pq}) = \begin{cases} \sigma_\nu^2 + \sigma_e^2, & i = p, j = q \\ \sigma_\nu^2, & i = p, j \neq q \\ 0, & i \neq p \end{cases} \tag{5.20}$$

분산성분 모형 (5.18)과 (5.19)는 county내의 sample segment에서 옥수수와 콩의 상관구조를 전적으로 정의하지는 못한다. 그러나 이 연구에서는 사용된 자료가 다

변량 모형을 적용하였을 경우 추정의 정도를 개선하지 못했기 때문에 단순히 일변량의 경우로 제한하였다.

i번째 county에서 옥수수(또는 콩) 재배면적의 표본 평균은 $\bar{y}_{i\cdot} = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$로 주어지고, (5.18)식과 (5.19)식을 이용하여 표현하면 다음과 같다.

$$\bar{y}_{i\cdot} = \beta_0 + \beta_1 \bar{x}_{1i\cdot} + \beta_2 \bar{x}_{2i\cdot} + \nu_i + \bar{e}_{i\cdot} \tag{5.21}$$

여기에서 $\bar{x}_{1i\cdot} = \frac{\sum_{j=1}^{n_i} x_{1ij}}{n_i}$, $\bar{x}_{2i\cdot} = \frac{\sum_{j=1}^{n_i} x_{2ij}}{n_i}$, $\bar{e}_{i\cdot} = \frac{\sum_{j=1}^{n_i} e_{ij}}{n_i}$로 주어진다.

모평균을 y_i 라 표현한다면,

$$y_i = \beta_0 + \beta_1 \bar{x}_{1i(p)} + \beta_2 \bar{x}_{2i(p)} + \nu_i \ , \tag{5.22}$$

여기에서 $\bar{x}_{1i(p)} = \frac{\sum_{j=1}^{N_i} x_{1ij}}{N_i}$, $\bar{x}_{2i(p)} = \frac{\sum_{j=1}^{N_i} x_{2ij}}{N_i}$ 로 주어지고, $N_i =$ i번째 county에서 segment수의 총합을 나타낸다. 위성자료로부터 $\bar{x}_{1i(p)}$, $\bar{x}_{2i(p)}$ 가 계산된다.

(5.22)식을 기반으로 평균 농작물 재배면적에 관한 추정 문제를 다루고자 한다. 유한 모집단 모형에서 i번째 county에서 옥수수(또는 콩)의 평균 재배면적은 $\bar{Y}_{i\cdot} = \frac{\sum_{j=1}^{N_i} Y_{ij}}{N_i}$이고, (5.22)식의 y_i 는 표본 추출률이 작을 때 $\bar{Y}_{i\cdot}$에 대한 예측변수를 나타낸다.

(5.18) ~ (5.20)식을 행렬형식으로 표현하면, $Y_i = (y_{i1}, y_{i2}, \cdots, y_{in_i})^t$,

$Y=(\ Y_1{}^t,\ Y_2{}^t,\cdots,\ Y_T{}^t)$ 로부터

$$Y=X\beta+u\ , \tag{5.23}$$

여기에서 y_{ij} 에 대응되는 X의 행은 $x_{ij}=(1,x_{1ij},x_{2ij})$, $\beta=(\beta_0,\beta_1,\beta_2)^t$, u 의 공분산 행렬은

$$E(uu^t)=V=block\ diag(V_1,\ V_2,\cdots,\ V_T), \tag{5.24}$$

$$V_i=J_i\ \sigma_\nu{}^2+I_i\ \sigma_e{}^2\ , \tag{5.25}$$

여기에서 J_i는 order가 n_i 이며 모든 원소가 1인 정방행렬이고, I_i는 order가 n_i 인 단위행렬을 나타낸다.

위의 (5.22)식을 행렬로 표현하면 다음과 같다.

$$y_i=\ \bar{x}_{i(p)}\beta+\nu_i\ , \tag{5.26}$$

여기에서 $\bar{x}_{i(p)}=\dfrac{\sum_{j=1}^{N_i}x_{ij}}{N_i}=(1,\ \bar{x}_{1i(p)},\ \bar{x}_{2i(p)})$로 주어진다.

5.2.3 모수 추정

앞에서 언급한 (5.18), (5.19), (5.20)을 충족한다는 가정하에서 모수에 대한 추정을 고려해 보자. 만약에 u_{ij} 가 가지이면, ν_i 의 best predictor는 $E(\nu_i\mid\bar{u}_{i\cdot})$ 가 되고, 여기에서 $\bar{u}_{i\cdot}=\dfrac{\sum_{j=1}^{n_i}u_{ij}}{n_i}$ 이다.

(5.18)과 (5.19)의 가정하에서 ν_i 와 $\bar{u}_{i\cdot}$는 다음과 같은 이변량 정규분포를 따

르게 되고,

$$(\nu_i,\ \bar{u}_{i\cdot}) \sim N\left(0,\ \begin{pmatrix} \sigma_\nu^2 & \sigma_\nu^2 \\ \sigma_\nu^2 & \sigma_\nu^2 + \frac{\sigma_e^2}{n_i} \end{pmatrix}\right),$$

주변확률 분포로부터 $E(\nu_i \mid \bar{u}_{i\cdot}) = \bar{u}_{i\cdot} \dfrac{\sigma_\nu^2}{\sigma_\nu^2 + \frac{\sigma_e^2}{n_i}} = \bar{u}_{i\cdot}\, g_i$ 가 주어진다. 여기에서 $g_i = m_i^{-1}\,\sigma_\nu^2$ 이고 $m_i = (\sigma_\nu^2 + \frac{\sigma_e^2}{n_i})$ 이다.

따라서 위의 관계로부터 다음의 결과를 얻을 수 있다.

$$E(\nu_i - \bar{u}_{i\cdot}\, g_i \mid \bar{u}_{i\cdot}) = 0\ ,$$

$$E\{(\nu_i - \bar{u}_{i\cdot}\, g_i)^2\} = \sigma_\nu^2 (1 - g_i)$$

$$= n_i^{-1}\,\sigma_e^2 - n_i^{-2}\,\sigma_e^2\, m_i^{-1}\,\sigma_e^2 \qquad (5.27)$$

여기에서 σ_e^2 과 σ_ν^2은 미지인 값이며, 만약에 σ_e^2, σ_ν^2이 기지이면 β의 GLS 추정량은 다음과 같이 주어진다.

$$\hat{\beta} = (X^t V^{-1} X)^{-1} X^t V^{-1} Y \qquad (5.28)$$

이때 i번째 county의 효과 ν_i 는 다음 식으로 추정된다.

$$\tilde{\nu}_i = \tilde{u}_{i\cdot}\, g_i\ , \qquad (5.29)$$

단, $\tilde{u}_{i\cdot} = \dfrac{\sum_{j=1}^{n_i} \tilde{u}_{ij}}{n_i}$, $\tilde{u}_{ij} = y_{ij} - x_{ij}\hat{\beta}$ 로 주어진다. 여기에 대응되는 y_i (i번째 county의 평균 재배면적) BLUP 추정량은 다음의 (5.30)식으로 주어지고,

$$\tilde{y}_i = \bar{x}_{i(p)}\tilde{\beta} + \tilde{\nu}_i \tag{5.30}$$

(5.30)식으로부터 다음의 결과를 얻을 수 있다.

$$E\{(\tilde{y}_i - y_i)^2\} = \sigma_\nu^2(1-g_i) + c_i V(\tilde{\beta})\, c_i^t \tag{5.31}$$

단, $V(\tilde{\beta}) = (X^t V^{-1} X)^{-1}$,

$$c_i = \bar{x}_{i(p)} - g_i \bar{x}_{i\cdot}, \quad \bar{x}_{i\cdot} = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i} = (1, \bar{x}_{1i\cdot}, \bar{x}_{2i\cdot})$$

$$x_{ij} = (1, x_{1ij}, x_{2ij}),$$

$$\bar{x}_{i(p)} = \frac{\sum_{j=1}^{N_i} x_{ij}}{N_i} = (1, \bar{x}_{1i(p)}, \bar{x}_{2i(p)})$$

(5.31)식은 (5.27)식 보다 $c_i V(\tilde{\beta})\, c_i^t$ 만큼 큰 양을 나타낸다.

유한 모집단에서 i번째 county에서 재배면적의 모평균에 대한 추정은 다음식을 이용한다.

$$\frac{\sum_{j=1}^{n_i} y_{ij} + \sum_{j=n_i+1}^{N_i} (x_{ij}\tilde{\beta} + \tilde{\nu}_i)}{N_i} \tag{5.32}$$

이 추정식은 표본 추출률이 작을 경우에는 (5.30)식과 근사적으로 같다. 이러한 사실을 근거로 해서 (5.30)식을 이용하여 문제에 접근하였다.

대부분의 문제에서는 σ_ν^2과 σ_e^2은 미지이므로 추정되어야 한다. β 의 GLS 추정량 $\tilde{\beta} = (X^t V^{-1} X)^{-1} X^t V^{-1} Y$ 과 $\tilde{y}_i = \bar{x}_{i(p)}\tilde{\beta} + \tilde{\nu}_i$ 는 σ_ν^2과 σ_e^2이 미지이므로 추정되어야 한다.

식(5.18)에서 가정한 모형으로부터

$$\hat{\sigma}_e^2 = \hat{e}^t \hat{e} \Big(\sum_{i=1}^{T} (n_i - 1) - 2 \Big), \tag{5.33}$$

여기에서 $\hat{e}^t \hat{e}$ 는 $(y_{ij} - \bar{y}_{i.})^2$ 의 합이며, $n_i > 1$일때 (5.18)과 (5.19)의 가정하에서 $\hat{\sigma}_e^2$은 σ_e^2에 대해서 불편이고, $d_e \frac{\hat{\sigma}_e^2}{\sigma_e^2} \sim \chi^2(d_e)$ 인 관계가 성립하므로 $d_e = \sum_{i=1}^{T}(n_i - 1) - 2$ 로 주어진다.

county 효과 $\hat{\sigma}_v^2$ 은 i county 에 대해서 다음의 (5.34)식의 잔차를 고려하여 구해진다.

$$\breve{u}_{i.} = \bar{y}_{i.} - \bar{x}_{i.} (X^t X)^{-1} X^t Y \ , \tag{5.34}$$

여기에서

$$E(\breve{u}_{i.}^{\,2}) = b_i \sigma_v^2 + d_i \sigma_e^2 \ , \tag{5.35}$$

단, $b_i = 1 - 2 n_i \bar{x}_{i.} (X^t X)^{-1} X^t \bar{x}_{i.}^{\,t}$

$$+ \bar{x}_{i.} (X^t X)^{-1} \Big(\sum_{j=1}^{T} n_j^2 \bar{x}_{j.}^{\,t} \bar{x}_{j.} \Big) (X^t X)^{-1} \bar{x}_{i.}^{\,t} \ ,$$

$$d_i = n_i^{-1} \{ 1 - n_i \bar{x}_{i.} (X^t X)^{-1} \bar{x}_{i.}^{\,t} \}$$

county 들에 대한 잔차의 가중제곱합의 평균은 다음의 (5.36)식으로 주어지고,

$$\hat{m}_{..} = \frac{\sum_{i=1}^{T} n_i \breve{u}_i^2}{\sum_{i=1}^{T} n_i b_i} \ , \tag{5.36}$$

위의 (5.36)식의 기대값은 다음식과 같이 표현된다.

$$E(\hat{m}\,..) = E\left(\frac{\sum_{i=1}^{T} n_i \breve{u}_i^{\,2}}{\sum_{i=1}^{T} n_i b_i}\right)$$

$$= m\,..$$

$$= \sigma_\nu^{\,2} + c\,\sigma_e^{\,2}$$

$$\text{단, } c = \frac{\sum_{i=1}^{T} n_i d_i}{\sum_{i=1}^{T} n_i b_i}$$

(5.18)과 (5.19)의 가정하에서 $\hat{m}\,..$ 와 $\hat{\sigma}_e^{\,2}$은 서로 독립이므로 $\sigma_\nu^{\,2}$의 추정량은 다음의 결과를 이용한다.

$$\hat{\sigma}_\nu^{\,2} = \max\left\{\hat{m}\,.. - c\,\hat{\sigma}_e^{\,2}\ ,\ 0\right\} \tag{5.37}$$

이상의 추정에 근거하면 $g_i = \dfrac{\sigma_\nu^{\,2}}{\sigma_\nu^{\,2} + \dfrac{\sigma_e^{\,2}}{n_i}}$ 의 추정식 $\tilde{g}_i$ 는 다음의 (5.38)식과 같고,

$$\tilde{g}_i = \frac{\hat{\sigma}_\nu^{\,2}}{\hat{\sigma}_\nu^{\,2} + \dfrac{\hat{\sigma}_e^{\,2}}{n_i}} \tag{5.38}$$

따라서 (5.26)식 $y_i = \bar{x}_{i(p)}\beta + \nu_i$, $\bar{x}_{i(p)} = \dfrac{\sum_{j=1}^{N_i} x_{ij}}{N_i} = (1,\ \bar{x}_{1i(p)},\ \bar{x}_{2i(p)})$는 다음 식으로 추정될 수 있다.

$$\hat{y}_i = \bar{x}_{i(p)}\hat{\beta} + \hat{u}_{i\cdot}\ \hat{g}_{i\cdot}\ , \tag{5.39}$$

여기에서 $\hat{\beta}$ 은 β 의 GLS 추정량인 $\tilde{\beta}$ 에서 V를 $\hat{V}$ 로 대체하고 다음과 같이 구할 수 있다.

$$\hat{\beta} = (X^t \hat{V}^{-1} X)^{-1} X^t \hat{V}^{-1} Y ,$$

단, $\hat{V}(= \widehat{E(u u^t)}) = block\ diag(\hat{V}_1, \hat{V}_2, \cdots, \hat{V}_T)$, $\hat{V}_i = J_i \hat{\sigma}_\nu^2 + I_i \hat{\sigma}_e^2$,

$\hat{u}_{i\cdot} = \bar{y}_{i\cdot} - \bar{x}_{i\cdot} \hat{\beta}$, $\hat{g}_i = g_i$ 에 대해서 불편인 추정량이다.

추정오차 $y_i - \hat{y}_i$ 의 분산에 대한 추정식은 다음의 결과를 이용한다(다변량의 경우에는 Fuller and Harter(1987)를 참조).

$$\begin{aligned} \hat{y}_i &= \bar{x}_{i(p)} \hat{\beta} + \hat{u}_{i\cdot} \hat{g}_i \\ &= \bar{x}_{i(p)} \hat{\beta} + (\bar{y}_{i\cdot} - \bar{x}_{i\cdot} \hat{\beta}) \hat{g}_i \quad , \end{aligned} \tag{5.40}$$

여기에서 $\hat{g}_i = 1 - \hat{h}_i$,

$$\hat{h}_i = \{ \hat{m}_i + \hat{k}_i + (n_i^{-1} - c)^2 \hat{w}_i \}^{-1} \{ n_i^{-1} \hat{\sigma}_e^2 + (n_i^{-1} - c)\ n_i^{-1} \hat{w}_i \},$$

단, $\hat{m}_i = \hat{m}_{\cdot\cdot} + (n_i^{-1} - c)\ \hat{\sigma}_e^2$,

$$\hat{w}_i = 2\ d_e^{-1}\ \hat{m}_i^{-1}\ \hat{\sigma}_e^4 ,$$

$$\hat{k}_i = 2\ \hat{\sigma}_e^2 (\ddot{\sigma}_{ff} + n_i^{-1})^{-1} (\sum_{j=1}^{T} n_j b_j)^{-2} (\sum_{j=1}^{T} n_j^2 b_j (\ddot{\sigma}_{ff} + n_j^{-1})^2) ,$$

단, $\ddot{\sigma}_{ff} = \max \{ 0 ,\ (T-5)^{-1} (T-3)\ \hat{\sigma}_e^{-2}\ \hat{m}_{\cdot\cdot} - c \} ,$

$$b_j = 1 - 2 n_j \bar{x}_{j\cdot} (X^t X)^{-1} X^t\ \bar{x}_{j\cdot}^{\ t}$$

$$+ \bar{x}_{j\cdot}(X^tX)^{-1}\Big(\sum_{i=1}^{T} n_i^2 \ \bar{x}_{i\cdot}{}^t \bar{x}_{i\cdot}\Big)(X^tX)\ \bar{x}_{j\cdot}{}^t ,$$

$$c = \frac{\sum_{i=1}^{T} n_i d_i}{\sum_{i=1}^{T} n_i b_i} ,$$

$$d_i = n_i^{-1}(1 - n_i \bar{x}_{i\cdot}(X^tX)^{-1}\ \bar{x}_{i\cdot}{}^t) ,$$

$$d_e = 22$$

(5.40)식의 결과를 이용하여 추정오차의 분산의 추정식을 계산하면 다음과 같이 주어진다.

$$\widehat{Var}(\hat{y}_i - y_i) = n_i^{-1}\ \hat{\sigma}_e^2 - \hat{\phi}_i + \hat{c}_i \hat{V}(\hat{\beta})\ \hat{c}_i^t + \hat{h}_i^2 \hat{k}_i + d_e^{-1}\ \hat{r}_i^2 \hat{\phi}_i + d_e^{-1}\ \hat{r}_i^2 \hat{h}_i\ \hat{\sigma}_e^2,$$

단, $\hat{c}_i = \bar{x}_{i(p)} - \hat{g}_i \bar{x}_{i\cdot}$,

$$\hat{\phi}_i = (d_e+1)^{-1} d_e \ddot{\phi}_i - d_e^{-1}\ n_i^{-1}\ \hat{\sigma}_e^2 \hat{h}_i ,$$

$$\ddot{\phi}_i = n_i^{-2}(\hat{\sigma}_e^2 + (n_i^{-1} - c)\hat{w}_i)^2 (\hat{m}_i + \hat{k}_i + (n_i^{-1} - c)^2 \hat{w}_i)^{-1} ,$$

$$\hat{r}_i = 1 - (1 - n_i c)\hat{h}_i$$

5.2.4 추정 결과

i) $\hat{\sigma}_\nu^2 = \max\{\ \hat{m}_{\cdot\cdot} - c\ \hat{\sigma}_e^2\ ,\ 0\}$

$$\hat{\sigma}_e^2 = \hat{e}^t \hat{e}\Big(\sum_{i=1}^{T}(n_i - 1) - 2\Big) , \quad n_i > 1$$

ii) β 의 GLSE

$$\hat{\beta} = (X^t \hat{V}^{-1} X)^{-1} X^t \hat{V}^{-1} Y ,$$

단, $\hat{V} = block\ diag(\hat{V}_1, \hat{V}_2, \cdots, \hat{V}_T)$

$$\hat{V}_i = J_i \hat{\sigma}_v^2 + I_i \hat{\sigma}_e^2$$

iii) $\hat{y}_i = \bar{x}_{i(p)} \hat{\beta} + \hat{u}_i \cdot \hat{g}_i$

$= \bar{x}_{i(p)} \hat{\beta} + (\bar{y}_{i\cdot} - \bar{x}_{i\cdot} \hat{\beta}) \hat{g}_i$

옥수수의 경우

$$\hat{y}_{ij} = 51 + 0.329 x_{1ij} - 0.134 x_{2ij} ,$$

(25) (0.050) (0.056)

$$\hat{\sigma}_e^2 = 150 , \quad \hat{\sigma}_v^2 = 140$$

(45) (89)

콩의 경우

$$\hat{y}_{ij} = -16 + 0.028 x_{1ij} + 0.494 x_{2ij}$$

(29) (0.058) (0.065)

$$\hat{\sigma}_e^2 = 195 , \quad \hat{\sigma}_v^2 = 272$$

(59) (49)

$$c = 0.349$$

<표4> 옥수수의 예측 재배면적과 표준오차

County	Sample segment	Predicted hectares	standard error		
			Best predictor	Survey regression predictor	Sample mean
Cerro Gordo	1	122.2	9.6	13.7	30.5
Hamilton	1	126.3	9.5	12.9	30.5
Worth	1	106.2	9.3	12.4	30.5
Humboldt	2	108.0	8.1	9.7	21.5
Franklin	3	145.0	6.5	7.1	17.6
Pocahontas	3	112.6	6.6	7.2	17.6
Winnebago	3	112.4	6.6	7.2	17.6
Wright	3	122.1	6.7	7.3	17.6
Webster	4	115.8	5.8	6.1	15.2
Hancock	5	124.3	5.3	5.7	13.6
Kossuth	5	106.3	5.2	5.5	13.6
Hardin	5	143.6	5.7	6.1	13.6

* Best Predictor : $\hat{y_i} = \bar{x}_{i(p)}\hat{\beta} + \hat{u}_i \cdot \hat{g_i}$

* Survey Regresson Predictor : $y_i = \bar{y}_{i\cdot} + (\bar{x}_{i(p)} - \bar{x}_{i\cdot})\hat{\beta}$

* 표본평균의 추정 표준오차는 county내의 mean square를 county내의 segment의 수로 나누어 root를 취한 값임.

<표5> 콩의 예측 재배면적과 표준오차

County	Sample segment	Predicted hectares	standard error		
			Best predictor	Survey regression predictor	Sample mean
Cerro Gordo	1	77.8	12.0	15.6	29.1
Hamilton	1	94.8	11.8	14.8	29.1
Worth	1	86.9	11.5	14.2	29.1
Humboldt	2	79.7	9.7	11.1	20.6
Franklin	3	65.2	7.6	8.1	16.8
Pocahontas	3	113.8	7.7	8.2	16.8
Winnebago	3	98.5	7.7	8.3	16.8
Wright	3	112.8	7.8	8.4	16.8
Webster	4	109.6	6.7	7.0	14.6
Hancock	5	101.0	6.2	6.5	13.0
Kossuth	5	119.9	6.1	6.3	13.0
Hardin	5	74.9	6.6	6.9	13.0

이상의 추정결과를 요약하면 다음과 같이 정리할 수 있다. 첫째, sample segment가 증가할수록 표본평균의 표준오차는 감소하는 경향을 볼 수 있다. 둘째, 표본평균의 표준오차는 survey regression predictor의 표준오차보다 상당히 크다는 사실을 확인할 수 있다($\frac{best\ predictor의\ 표준오차}{survey\ regression\ predictor의\ 표준오차} = 0.77 \sim 0.97$). 또한 sample segment의 수가 3 이하일 때 Best Predictor는 작은 표준오차를 가지며, Predictor의 정확도는 sample segment의 수가 한 county에서 3에서 4 또는 5정도일때 정확도가 좋게 나타났다.

Survey Regression Predictor는 12개의 county의 평균 농작재배면적에 대해서 unbiased되어있고, 상대적으로 작은 분산을 갖고 있다. 따라서 Survey Regression Predictor는 전체 지역에 대해서 적절한 Predictor라고 볼수 있다. 각각의 county에 대한 예측은 적당한 가중합을 이용하여 전지역에 대하여 불편인 특성을 갖는 Survey Regression Predictor와 같게 되도록 수정하는것도 생각해 볼 문제이다.

결론적으로 보조 변수로써 위성관측 자료를 이용한 분산성분 모형 적용은 소지역에서 농작 재배면적을 예측하는데 유용한 결과를 제공한다는 사실을 확인할 수 있다.

5.3 암시적/명시적 모형을 이용한 소지역 추정 사례(III)

5.3.1 서 론

갤럽은 미국 전역의 알콜중독과 마약복용률을 추정하기 위해 주(State) 단위의 광범위한 가구조사를 실시하고 있다. 추정값은 주(State) 단위에서는 비교적 신뢰할만한 수준이나 sub-state 그룹에서의 추정은 표본이 불균형적으로 배정되거나 표본크기가 작을 경우에는 추정의 신뢰도가 상당히 떨어진다. 주(State) 단위로 조사된 조사 자료를 기반으로 sub-state에서의 추정을 살펴보며, 비교 추정량으로 직접추정량, 합성추정량, 복합추정량, 경험적 베이즈 추정량이 제시된다.

$n_i (i = 1, 2, \cdots, I)$를 i번째 계획구역에 할당된 표본크기로 하자($n = \sum_{i=1}^{I} n_i$). 각각의 계획구역에서의 표본은 RDD(random digit dialing)전화조사로 독립적으로 추출된다. 표본이 관측된 후 각각의 구역은 K개의 인구통계적 그룹으로 사후 층화된다. 이러한 그룹들은 성별(남, 여), 나이(18-24, 25-44, 45-64, 65세 이상)로 교차 분류되며 총 $K = 2 \times 4 = 8$개의 그룹으로 분류된다.

J_i 를 i번째 계획구역에 있는 county 개수라 하고, n_{ijk}를 i번째 계획구역에

속해있는 j번째 county에서 k번째 인구통계적 그룹에 있는 관측값의 개수라 하자. 또한 S_{ij}는 i번째 계획구역 내의 j번째 county에 있는 인구통계적 그룹의 set을 나타내고, y_{ijkl}는 i번째 계획구역에 속해있는 j번째 county의 k번째 인구통계적 그룹에 대한 l번째 관측값(0 or 1), w_{ijkl}는 표본조사로부터 얻은 표본추출가중치라 하자($i=1,2,\cdots,I$; $j=1,2,\cdots,J_i$; $k \in S_{ij}$; $l=1,2,\cdots,n_{ijk}$). 이때 i번째 계획구역내에 있는 j번째 county에 대한 알콜중독 또는 마약복용률을 나타내는 π_{ij}를 추정하고자 한다.

5.3.2 직접추정량과 합성추정량

j번째 county에 대한 π_{ij} 의 직접추정값은 다음 (5.41)식으로 계산하고,

$$\widehat{\pi}_{ij}^{D} = \frac{\sum_{k \in S_{ij}} \sum_{l=1}^{n_{ijk}} w_{ijkl}\, y_{ijkl}}{\sum_{k \in S_{ij}} \sum_{l=1}^{n_{ijk}} w_{ijkl}} \tag{5.41}$$

k번째 인구통계적 그룹에 대한 π_{ik}의 직접추정값은 다음 (5.42)식으로 계산된다.

$$\widehat{\pi}_{ik}^{D} = \frac{\sum_{j:\, k \in S_{ij}} \sum_{l=1}^{n_{ijk}} w_{ijkl}\, y_{ijkl}}{\sum_{j:\, k \in S_{ij}} \sum_{l=1}^{n_{ijk}} w_{ijkl}} \tag{5.42}$$

여기에서 $j: k \in S_{ij}$ 는 인구통계적 그룹 k가 관측된 모든 county j에 대한 합을 의미한다.

이 추정량은 하나의 county로부터 이용할 수 있는 표본의 크기가 매우 작을 경우 추정값의 신뢰도는 매우 떨어진다. 또한, 하나의 county내에 있는 모든 관측점이

0이라면, 추정값도 0이 되고, 추정값의 표준오차도 0이 되어 추정의 신뢰도가 떨어진다. 따라서 직접추정량에서는 이러한 문제가를 개선되어야 할 필요가 있다.

합성추정량은 전화조사 자료와 U.S 센서스 국으로부터 획득한 보조자료의 관계를 이용하여 추정한다. county수준에서의 과거 알콜중독률을 이용한 합성추정량은 다음과 같다.

$$\widehat{\pi}_{ij}^{S1} = \sum_{k=1}^{K} a_{ijk} \quad \widehat{\pi}_{k}^{D} \tag{5.43}$$

여기에서 $\widehat{\pi}_{k}^{D}$ = k번째 인구통계 그룹에 대한 알콜중독률의 직접조사추정량, a_{ijk} = i번째 계획구역의 j번째 county의 k번째 인구통계적 그룹에 속해 있는 개체들의 비율이다(최근의 센서스 추정치로부터 얻음). $\widehat{\pi}_{ij}^{S1}$에서 나타나 있듯이 k번째 인구통계적 그룹에 대한 알콜중독률과 마약복용률은 모든 county들에 대해서 동일한 것으로 가정되었다.

좀더 덜 제한적인 알콜중독과 마약복용률에 대한 합성추정량은 다음과 같다.

$$\widehat{\pi}_{ij}^{S2} = \sum_{k=1}^{K} a_{ijk} \quad \widehat{\pi}_{ik}^{D} \tag{5.44}$$

여기에서 $\widehat{\pi}_{ik}^{D}$는 i번째 계획구역에 있는 k번째 인구통계적 그룹에 대한 알콜중독 또는 마약복용률 π_{ik}의 직접조사추정량을 나타낸다. $\widehat{\pi}_{ij}^{S2}$에서는 k번째 그룹에 대한 알콜중독률과 마약복용률은 하나의 계획구역에 있는 모든 county에 대해서 동일한 것으로 가정되었고, 이러한 가정은 $\widehat{\pi}_{ij}^{S1}$에서 보다는 합리적이고 덜 제한적인 가정이라 할 수 있다.

5.3.3 복합 추정량

직접조사추정량과 합성추정량의 절충이 복합추정량이다. 여기서 제안하는 복합추정량은 다음의 항등식에 근거한다.

$$\pi_{ij} = \sum_{k \in S_{ij}} a_{ijk} \pi_{ijk} + \sum_{k \notin S_{ij}} a_{ijk} \pi_{ijk} \tag{5.45}$$

여기에서 π_{ijk} = 알콜중독 또는 마약복용률, a_{ijk} = i번째 계획구역의 j번째 county에 있는 k번째 인구통계적 그룹에 속해있는 개체들의 비율을 나타낸다.

π_{ij}의 단순복합추정량은 $k \in S_{ij}$에 대해서 π_{ijk}는 $\widehat{\pi_{ijk}}^{D}$로, $k \notin S_{ij}$에 대해서 π_{ijk}는 $\widehat{\pi_{ik}}^{D}$로 추정되며 다음과 같다.

$$\widehat{\pi_{ij}}^{C} = \sum_{k \in S_{ij}} a_{ijk} \ \widehat{\pi_{ijk}}^{D} + \sum_{k \notin S_{ij}} a_{ijk} \ \widehat{\pi_{ik}}^{D} \tag{5.46}$$

(5.46)식에서 π_{ijk} ($k \in S_{ij}$)는 소표본에서 추정하므로 추정의 정확도가 떨어지게 되어, 따라서 최근의 보조정보를 이용하여 추정한다면 $\widehat{\pi_{ijk}}^{D}$ 에 대한 개선의 여지가 남아 있다. 이를 위하여 π_{ij}의 경험적 베이즈 추정량을 제안한다.

5.3.4 경험적 베이즈(EB) 추정량

경험적 베이즈 추정량을 산출하기에 앞서 다음을 가정하기로 한다.

i) π_{ijk} 들이 주어진 상태에서 y_{ijk}들은 서로 uncorrelate 되어 있고, 아래 조건을 만족한다.

$$E(y_{ijkl} \mid \pi_{ijk}) = \pi_{ijk} \ ,$$

$$Var(y_{ijkl} \mid \pi_{ijk}) = \pi_{ijk}(1 - \pi_{ijk}).$$

ii) π_{ijk} 들은 서로 uncorrelate 되어 있으며, 다음 식을 만족한다.

$$E(\pi_{ijk}) = \mu_{ik} ,$$

$$Var(\pi_{ijk}) = d\ \mu_{ik}^2 .$$

만약 $\pi_{ijk} \sim U(0, 2\mu_{ik})$ 라면, 위에서 $d = \frac{1}{3}$. 즉, $\widehat{\pi}_{ij}^{S2}$에서의 가정 ($\pi_{ijk} = \mu_{ik}$)과는 달리, 특정한 인구통계적 그룹에 대해서 한 구역 내의 county들 간의 비율의 변동이 반영된다.

첫 번째 가정은 π_{ijk}가 주어진 상태에서 $\widehat{\pi}_{ijk}^{D}$는 서로 uncorrelate 되어 있음을 의미한다. 이때,

$$E(\ \widehat{\pi}_{ijk}^{D} \mid \pi_{ijk}) = \pi_{ijk} ,$$

$$Var(\ \widehat{\pi}_{ijk}^{D} \mid \pi_{ijk}) = c_{ijk}\ \pi_{ijk}(1-\pi_{ijk}) ,$$

$$\text{단, } c_{ijk} = \frac{\sum_{l=1}^{n_{ijk}} w_{ijkl}^2}{(\sum_{l=1}^{n_{ijk}} w_{ijkl})^2} .$$

손실함수로 제곱오차손실함수(squared error loss function)가 사용될 경우, π_{ij}의 선형 베이즈 추정량(linear Bayes estimator)은 아래 (5.47)식과 같이 주어진다.

$$\widehat{\pi}_{ij}^{B} = \sum_{k \in S_{ij}} a_{ijk}(B_{ijk}\ \widehat{\pi}_{ijk}^{D} + (1-B_{ijk})\mu_{ik}) + \sum_{k \notin S_{ij}} a_{ijk}\mu_{ik} , \quad (5.47)$$

$$\text{단, } B_{ijk} = \frac{d\ \mu_{ik}^2}{d\ \mu_{ik}^2 + c_{ijk}(\mu_{ik} - (d+1)\ \mu_{ik}^2)}$$

위의 베이즈 추정량은 미지인 모수 μ_{ik}를 포함하고 있기 때문에 μ_{ik}가 먼저 추

정되어야만 한다. μ_{ik}가 $\widehat{\mu_{ik}}$ (= $\widehat{\pi_{ik}}^{D}$)로 대체되어 얻어진다면, π_{ij}의 경험적 베이즈 추정량 $\widehat{\pi_{ij}}^{EB}$는 다음과 같이 주어질 수 있다.

$$\widehat{\pi_{ij}}^{EB} = \sum_{k \in S_{ij}} a_{ijk}(\widehat{B_{ijk}}\ \widehat{\pi_{ijk}}^{D} + (1 - \widehat{B_{ijk}})\ \widehat{\mu_{ik}}) + \sum_{k \notin S_{ij}} a_{ijk}\ \widehat{\mu_{ik}}\ , \quad (5.48)$$

$$\text{단,}\quad \widehat{B_{ijk}} = \frac{d\ \widehat{\mu_{ik}}^{2}}{d\ \widehat{\mu_{ik}}^{2} + c_{ijk}(\widehat{\mu_{ik}} - (d+1)\ \widehat{\mu_{ik}}^{2})}$$

가중치 또는 축소인자(shrinkage factors)로 불리우는 $\widehat{B_{ijk}}$는 $\widehat{\pi_{ijk}}$의 분산에 대한 π_{ijk}의 분산의 비이고, $\widehat{\mu_{ik}}$는 $\widehat{\pi_{ik}}^{D}$로 대체되어 구해진다.

베이즈 추정량의 MSE는 $MSE(\ \widehat{\pi_{ij}}^{B}) = E(\ \widehat{\pi_{ij}}^{B} - \pi_{ij})^2$로 구해지고 i)과 ii)의 가정하에서 다음과 같이 계산된다.

$$\begin{aligned} MSE(\ \widehat{\pi_{ij}}^{B}) &= Var(\ \widehat{\pi_{ij}}^{B} - \pi_{ij}) \\ &= Var(\ \widehat{\pi_{ij}}^{B}) + Var(\pi_{ij}) - 2\,Cov(\ \widehat{\pi_{ij}}^{B}, \pi_{ij}) \\ &= Var(\pi_{ij}) - Var(\ \widehat{\pi_{ij}}^{B}) \\ &= d\Big\{ \sum_{k \in S_{ij}} a_{ijk}^{2}(1 - B_{ijk})\ \mu_{ik}^{2} + \sum_{k \notin S_{ij}} a_{ijk}^{2}\ \mu_{ik}^{2} \Big\} \end{aligned}$$

경험적 베이즈 추정량의 MSE는 다음식으로 주어진다.

$$MSE(\ \widehat{\pi_{ij}}^{EB}) = MSE(\ \widehat{\pi_{ij}}^{B}) + E(\ \widehat{\pi_{ij}}^{EB} - \ \widehat{\pi_{ij}}^{B})^2 \quad (5.49)$$

(5.49)식은 미지인 모수 μ_{ik}를 포함하고 있기 때문에 추정되어야 한다.

첫 번째 항 $MSE(\ \widehat{\pi_{ij}}^{B})$는 다음 식으로 추정된다(Jiang et al.(1998)).

$$mse_J(\widehat{\pi}_{ij}^{B}) = mse(\widehat{\pi}_{ij}^{B}) - \frac{J_i-1}{J_i}\sum_{u=1}^{J_i}\{mse_{(-u)}(\widehat{\pi}_{ij}^{B}) - mse(\widehat{\pi}_{ij}^{B})\},$$

단, $$mse(\widehat{\pi}_{ij}^{B}) = d\Big\{\sum_{k\in S_{ij}} a_{ijk}^{2}(1-B_{ijk})\,\mu_{ik}^{2} + \sum_{k\notin S_{ij}} a_{ijk}^{2}\,\widehat{\mu_{ik}^{2}}\Big\},$$

$$mse_{(-u)}(\widehat{\pi}_{ij}^{B}) = d\Big\{\sum_{k\in S_{ij}} a_{ijk}^{2}(1-\widehat{B_{ijk(-u)}})\,\widehat{\mu_{ik(-u)}^{2}} + \sum_{k\notin S_{ij}} a_{ijk}^{2}\,\widehat{\mu_{ik(-u)}^{2}}\Big\},$$

단, $$\widehat{\mu_{ik(-u)}} = \frac{\sum_{j\neq u}^{J_i}\sum_{l=1}^{n_{ijk}} w_{ijkl}\,y_{ijkl}}{\sum_{j\neq u}^{J_i}\sum_{l=1}^{n_{ijk}} w_{ijkl}},$$

$$\widehat{B_{ijk(-u)}} = \frac{d\,\widehat{\mu_{ik(-u)}^{2}}}{d\,\widehat{\mu_{ik(-u)}^{2}} + c_{ijk}\{\widehat{\mu_{ik(-u)}} - (d+1)\,\widehat{\mu_{ik(-u)}^{2}}\}}.$$

두 번째 항 $E(\widehat{\pi}_{ij}^{EB} - \widehat{\pi}_{ij}^{B})^2$ 은 다음의 Jackknife 추정법으로 추정될 수 있다(Shao and Tu(1995)).

$$E_J(\widehat{\pi}_{ij}^{EB} - \widehat{\pi}_{ij}^{B})^2 = \frac{J_i-1}{J_i}\sum_{u=1}^{J_i}(\widehat{\pi}_{ij(-u)}^{EB} - \widehat{\pi}_{ij}^{EB})^2,$$

여기에서 $$\widehat{\pi}_{ij(-u)}^{EB} = \sum_{k\in S_{ij}} a_{ijk}\{\widehat{B_{ijk(-u)}}\,\widehat{\pi}_{ijk}^{D} + (1-\widehat{B_{ijk(-u)}})\,\widehat{\mu_{ik(-u)}}\}$$

$$+ \sum_{k\notin S_{ij}} a_{ijk}\,\widehat{\mu_{ik(-u)}}$$

따라서 $MSE(\widehat{\pi}_{ij}^{EB})$ 의 추정량은 다음식으로 주어진다.

$$mse(\widehat{\pi}_{ij}^{EB}) = mse_J(\widehat{\pi}_{ij}^{B}) + E_J(\widehat{\pi}_{ij}^{EB} - \widehat{\pi}_{ij}^{B})^2 \qquad (5.50)$$

5.3.5 추정 결과

<표 6> 40개의 county에 대한 알콜중독률의 5개 추정량(%)

(Est.se=직접추정치의 표준오차, $\sqrt{Est.\ mse}$=EB추정치의 추정된 MSE의 제곱근)

	Estimator								
	Direct		Synthetic1	Synthetic2	Composite	Empirical Bayes		Sample Size	Number of Groups Observed in County
county	$\hat{\pi}_{ij}^{D}$	(Est.se)	$\hat{\pi}_{ij}^{S1}$	$\hat{\pi}_{ij}^{S2}$	$\hat{\pi}_{ij}^{C}$	$\hat{\pi}_{ij}^{EB}$	$\sqrt{Est.\ mse}$		
1	1.7	(2.4)	3.4	1.6	0.9	1.6	(0.33)	30	8
2	4.4	(2.0)	3.8	1.8	7.2	2.1	(0.35)	111	8
3	0.0	(0.0)	3.6	3.3	0.0	3.0	(0.85)	36	8
4	0.0	(0.0)	3.3	5.6	1.6	5.3	(1.79)	6	5
5	9.4	(4.8)	3.3	5.6	14.1	6.9	(1.78)	37	8
6	1.6	(1.1)	3.4	3.0	1.7	2.7	(0.67)	136	8
7	9.3	(5.8)	3.4	3.1	9.9	3.1	(0.81)	25	6
8	0.0	(0.0)	3.6	3.2	0.4	3.1	(0.84)	20	7
9	0.0	(0.0)	3.4	5.8	5.6	5.8	(1.93)	3	3
10	1.5	(1.3)	3.4	2.1	0.7	1.9	(0.54)	81	8
11	0.0	(0.0)	3.3	1.6	0.0	1.5	(0.33)	58	8
12	7.0	(6.8)	3.5	1.7	5.0	1.8	(0.35)	14	6
13	5.7	(3.8)	3.3	5.5	12.9	6.4	(1.75)	37	8
14	0.0	(0.0)	3.5	1.7	0.8	1.6	(0.33)	12	4
15	2.4	(1.4)	3.3	5.6	2.0	4.4	(1.56)	120	8
16	4.1	(3.5)	3.3	3.0	2.5	3.0	(0.77)	32	7
17	2.8	(2.4)	3.8	1.8	1.3	1.8	(0.37)	48	8
18	3.9	(1.1)	3.4	3.0	3.2	3.2	(0.60)	316	8
19	0.0	(0.0)	3.4	5.7	3.7	5.7	(1.95)	19	5
20	3.1	(3.9)	3.6	3.2	14.9	3.2	(0.82)	20	6
21	2.7	(1.6)	3.3	5.6	4.1	5.8	(1.50)	102	8
22	4.2	(1.8)	3.3	2.1	1.8	2.2	(0.42)	124	8
23	9.7	(2.7)	4.3	8.0	11.8	8.8	(2.11)	121	8
24	0.0	(0.0)	3.3	2.0	0.2	1.9	(0.54)	22	6
25	7.8	(4.7)	3.3	1.6	2.8	1.8	(0.33)	32	6
26	0.0	(0.0)	3.5	1.7	0.0	1.6	(0.37)	28	7
27	2.2	(1.8)	3.2	5.6	1.6	4.9	(1.74)	63	8
28	10.5	(13.7)	3.4	1.6	14.2	1.7	(0.35)	5	5
29	0.0	(0.0)	3.5	3.1	1.8	3.0	(0.81)	12	5
30	0.0	(0.0)	3.2	1.5	0.0	1.5	(0.33)	11	6
31	4.6	(3.2)	3.5	5.9	17.0	5.8	(1.87)	44	8
32	8.4	(3.8)	3.7	3.4	8.4	4.1	(0.84)	52	8
33	2.5	(1.3)	3.4	2.2	2.5	2.1	(0.50)	144	8
34	2.9	(2.4)	3.6	1.7	1.3	1.7	(0.35)	49	7
35	0.0	(0.0)	3.3	3.0	0.0	2.8	(0.77)	22	8
36	0.0	(0.0)	3.4	3.1	0.3	2.9	(0.82)	17	6
37	4.2	(4.0)	3.0	2.0	3.4	2.1	(0.54)	26	6
38	0.0	(0.0)	3.4	5.8	3.7	5.7	(1.97)	16	6
39	0.0	(0.0)	3.5	3.1	0.6	3.0	(0.81)	10	6
40	5.3	(1.9)	3.4	3.1	2.9	3.5	(0.69)	144	8

직접추정값은 변화가 심하고 경우에 따라서는 0으로 추정되며, 표준오차는 $\sqrt{\frac{\widehat{\pi}_{ij}^{D}(1-\widehat{\pi}_{ij}^{D})}{n_{ij}}}$로 추정될 수 있으며 또한 0으로 추정되는 경우가 있다. 합성추정값(S1)이 가장 안정적이며, 0으로 추정되는 값은 없고 약간의 변화성만 보인다. 반면, 복합추정값은 다른 추정값에 비해 훨씬 변화가 심하게 나타난다. 경험적인 베이즈 추정값들은 합성추정값(S2)과 매우 유사하며 추정된 MSE의 제곱근 값은 비교적 안정적으로 나타난다 (경험적 베이즈 추정량의 d 값은 $\frac{1}{3}$로 하였음). 특히 경험적 베이즈 추정량은 county의 표본크기가 작을 경우 매우 효과적임을 확인할 수 있다

<표7>은 <표6>에 나타나있는 대상 주의 모든 county들에 대한 알콜중독률에 대한 추정결과이다.

<표7> 40개의 county의 알콜중독률에 대한 5개 추정량들의 요약(%)

Estimator	Min.	Q1	Q2	Q3	Max.	Mean	S.D
Direct	0.0	0.0	2.2	4.3	10.5	2.8	3.2
Synthetic 1	3.0	3.3	3.4	3.5	4.3	3.5	0.2
Synthetic 2	1.5	1.8	3.0	4.4	8.0	3.2	1.7
Composite	0.0	0.4	1.7	4.6	17.5	3.7	4.8
EB	1.5	1.8	2.8	4.2	8.8	3.2	4.8

여기에서 합성추정값들과 복합추정값들의 평균이 직접추정값들의 평균보다 높게 나타나는데 이는 합성추정값과 복합추정값들은 직접추정값들에 비해 0으로 추정된 값들이 거의 없기 때문이다.

5.4 MCMC기법을 이용한 소지역 추정사례(Ⅳ)

5.4.1 서 론

소지역 추정문제를 해결하기 위한 다양한 방법의 연구가 최근에 들어 많은 사람들의 관심에 의해 진행되고 있는데, 특히 베이즈 방법은 모형을 통해 소지역들을 체계적으로 연결하여 소지역 추론을 할 수 있다는 점에서 광범위하게 이용되고 있다. 계층적 베이즈(HB) 방법과 경험적 베이즈(EB) 방법에 관한 응용 및 일반 이론에 관한 내용은 Datta and Ghosh(1991), Fay and Herriot(1979), Ghosh and Lahiri(1987, 1992), Prasad and Rao(1990), Stroud(1987, 1991) 등에 의해 연구되었으며, 주로 연속형 자료를 갖는 변량들에 관한 내용을 다루었다. 그러나 많은 경우 조사자료들은 이산형이거나 범주형 자료일 수 있으며, 이러한 경우에 위의 방법들을 직접적으로 이용하는 것은 한계가 있다.

이진 조사자료(binary survey data)로 경험적 베이즈(EB) 방법 또는 계층적 베이즈(HB) 방법을 이용하여 소지역의 비율을 추정하는 방법들은 Dempster and Tomberlin(1980), MacGibbon and Tomberlin(1989), Malec, Sedransk, and Tompkins(1993) 등에 의해 연구되었다. 또한, Nandram and Sedransk(1993)은 이단계 집락표본(two-stage cluster sample)의 이진 자료에 대한 베이즈 추정방법을 소개하였고, 뒤이어 Stroud(1994)에 의해 이단계 표본추출뿐만 아니라 단순임의추출, 층화추출, 집락추출된 조사자료에 대한 베이즈 추정연구가 소개되었다.

이진 모형(binary model)들은 이산형 자료와 연속형 자료를 동시에 통합하는 일반화된 선형모형(Generalized Linear Model)의 일부분으로 구분된다. Ghosh, Natarajan, Stroud, and Carlin(1998)은 이산형 또는 범주형으로 구분되는 조사자료(Survey data)를 계층모형에 적합시켜 베이즈 방법으로 소지역 추정을 해결하는 방법을 소개하였다.

베이즈 방법을 이용한 소지역 추정에는 주로 MCMC(Markov Chain Monte Carlo) 적분기법이 적용된다. 깁스 샘플러는 이러한 MCMC 방법의 일종이다. 여기

에 소개되는 자료는 1991년 캐나다 내의 지리적으로 구분된 15개 지역의 표본에 대해서 "당신이 근무하고 있는 작업장에서 건강상의 유해요소에 노출되어 나쁜 영향을 받은 적이 있는가?"라는 질문에 대한 응답 결과(① 그런적이 있다(yes) ② 그런적이 없다(no) ③ 건강상의 유해요소에 노출된 적이 없다(not exposed) ④ 해당 사항이 없다(not applicable or not stated))에 관한 조사자료이며, 이 조사자료들은 각각의 지역에 대해서 특성화 기준(연령별: 40세 이하, 41세 이상, 성별: 남, 여)에 의해 범주화 되었다. 이 조사자료를 기반으로 각각의 지역내에서 연령별, 성별 기준에 따라 질문 문항에 대한 응답비율을 추정하고자 하였다.

5.4.2 계층 모형

I개의 소지역이 있을 때, Y_{ij}를 i번째 소지역 내에 있는 j번째 단위에 대한 최소충분통계량이라 하자($i=1,2,\cdots,I$; $j=1,2,\cdots,n_i$). 이때 Y_{ij}는 서로 독립이며 다음과 같은 확률밀도함수를 갖는다고 가정한다.

$$f(y_{ij}|\theta_{ij},\phi_{ij}) = \exp\left\{\frac{y_{ij}\theta_{ij}-\psi(\theta_{ij})}{\phi_{ij}} + \rho(y_{ij};\phi_{ij})\right\} \qquad (5.51)$$

즉, Y_{ij}의 확률밀도함수를 지수족(Exponential Family)으로 가정하며, 여기에서 θ_{ij}는 정준모수(canonical parameter), ϕ_{ij} (>0)는 산포모수(dispersion parameter)를 나타내며, 산포모수 ϕ_{ij}는 기지로 가정한다. 통상적으로 (5.51)의 가정하에서는 $g(\theta_{ij}) = \psi'(\theta_{ij}) \quad = E(Y_{ij}\mid\theta_{ij})$ ($\equiv\mu_{ij}$)인 관계가 성립한다.

자연모수(natural parameter) θ_{ij}는 소지역간의 상관성을 고려하여 다음과 같은 구조로 모형화한다.

$$h(\theta_{ij}) = \quad \boldsymbol{x}_{ik}^{T}\boldsymbol{\beta} + u_i + \varepsilon_{ij}\,, \quad i=1,2,\cdots,I\,, \quad j=1,2,\cdots,n_i \qquad (5.52)$$

여기에서 h는 엄밀증가함수(strictly increasing function)이고, $\boldsymbol{x}_{ik}$는 기지인 $p\times 1$ 계획벡터(design vector)이며, $\boldsymbol{\beta}=(\beta_1, \beta_2, \cdots, \beta_p)^T$는 $p\times 1$ 회귀계수(random regression coefficient) 벡터, u_i는 랜덤효과들이다. 오차 항 ε_{ij}는 $\varepsilon_{ij} \sim iid\ N(0, \sigma^2)$이고, 랜덤 항 u_i는 $u_i \sim iid\ N(0, \sigma_u^{\,2})$이며, ε_{ij}와 u_i는 서로 독립이라 가정한다.

위의 (5.51)을 계층적 구조를 갖는 구조화된 모형으로 다음과 같이 표현할 수 있다. 아래의 표현식에서 $\boldsymbol{\theta}=(\theta_{11}, \theta_{12}, \cdots, \theta_{1n_1}, \cdots, \theta_{I1}, \theta_{I2}, \cdots, \theta_{In_I})^T$, $R_u = \sigma_u^{\,-2}$, $R=\sigma^{-2}$을 나타낸다.

(Ⅰ) $Y_{ij} \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{u}, R_u=r_u, R=r$

$$\overset{\text{ind}}{\sim}\ f(y_{ij}\mid\theta_{ij}, \phi_{ij}) = \exp\left\{\frac{y_{ij}\theta_{ij}-\phi(\theta_{ij})}{\phi_{ij}} + \rho(y_{ij};\phi_{ij})\right\}$$

(Ⅱ) $h(\theta_{ij}) \mid \boldsymbol{\beta}, \boldsymbol{u}, R_u=r_u, R=r$

$$\overset{\text{ind}}{\sim}\ N(\boldsymbol{x}_{ij}^{\,T}\boldsymbol{\beta} + u_i,\ r^{-1})$$

(Ⅲ) $u_i \mid \boldsymbol{\beta}, R_u=r_u, R=r$

$$\overset{\text{ind}}{\sim}\ N(0,\ r_u^{\,-1})$$

여기에서 모수 $\boldsymbol{\beta}, R_u=r_u, R=r$ 의 사전분포들은 다음과 같이 가정한다

(Ⅳ) $\boldsymbol{\beta} \sim uniform(R^p)$, $(p<m)$,

$$R_u \sim \text{gamma}\left(\frac{a}{2},\ \frac{b}{2}\right),$$

$R \sim$ gamma $\left(\frac{c}{2}, \frac{d}{2}\right)$ 이고, β, R_u, R은 서로 독립

(단, $f(z) \propto \exp(-\alpha z) z^{\beta-1} I_{(0,\infty)}(z)$일 때, $Z \sim$ gamma(α, β))

위에서 언급한 $g(\theta_{ij})$는 Y_{ij}의 조건부 분포의 중심부분으로써 우리의 관심을 여기에 집중시킨다. 즉, $\boldsymbol{y}=(y_{11}, y_{12}, \cdots, y_{1n_1}, \cdots, y_{I1}, y_{I2}, \cdots, y_{In_I})^T$가 주어진 상태에서 $g(\theta_{ij})$들의 평균, 분산, 공분산 계산에 관심이 있다.

사전분포 (Ⅳ)를 가정한 모형 (5.52)에서 $\boldsymbol{y}$ 에 대한 θ_{ij}의 조건부 결합 사후확률분포는 $a>0$, $c>0$, $\sum_i^m n_I - p + d > 0$, $m+b>0$ 일 때, 모든 y_{ij}와 ϕ_{ij} (>0)에 대해서 $\int_{\underline{\theta}_{ij}}^{\overline{\theta}_{ij}} \exp\left\{\frac{\theta y_{ij} - \psi(\theta)}{\phi_{ij}}\right\} h(\theta)\, d\theta < \infty$ 이고, 주어진 $\boldsymbol{y}$에 대한 θ_{ij}들의 결합 사후확률분포는 진분포(proper distribution)이다. 여기에서 θ_{ij}의 적분구간은 $\pm\infty$를 포함하는 개구간(open interval)이다.

일반화된 선형모형(GLM)에서 $Y_{ij}|p_{ij} \sim Bin(n_{ij}, p_{ij})$인 경우를 생각해 보자. 이때 $Y_{ij}|p_{ij} \propto \exp\left\{y_{ij} \log \frac{p_{ij}}{1-p_{ij}} + n_{ij} \log(1-p_{ij})\right\}$이며, 정준연결(canonical link)을 고려하면, $\theta_{ij} = \log(p_{ij}/1-p_{ij})$ 이고, $g(\theta_{ij}) = \psi'(\theta_{ij})/n_{ij} = \exp\{\theta_{ij}/(1+\exp(\theta_{ij}))\}$이 된다. $g(\theta_{ij})$의 평균, 분산, 공분산 등은 θ_{ij}의 분포로부터 계산되며, 이를 위한 조건부 사후확률분포들은 위의 (Ⅰ)~(Ⅳ)를 이용하여 계산하면 다음의 (a)~(e)와 같이 주어진다.

(a) $\boldsymbol{\beta}|\ \boldsymbol{\theta}, \boldsymbol{u}, r_u, r, \boldsymbol{y}$

$$\sim N_p\left(\frac{\sum_i \sum_j [h(\theta_{ij})\ \boldsymbol{x}_{ij} - \boldsymbol{x}_{ij}\ u_i]}{\sum_i \sum_j \boldsymbol{x}_{ij}\ \boldsymbol{x}_{ij}^T},\quad r^{-1}\left(\sum_i \sum_j \boldsymbol{x}_{ij}\ \boldsymbol{x}_{ij}^T\right)^{-1}\right)$$

(b) $u_i \mid \boldsymbol{\theta}, \boldsymbol{\beta}, r_u, r, \boldsymbol{y}$

$$\overset{ind}{\sim} N(rn_i + r_u)^{-1} r \sum_j [h(\theta_{ij}) - \boldsymbol{x}_{ij}^T \boldsymbol{\beta}], \quad (rn_i + r_u)^{-1})$$

(c) $R \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{u}, r_u, \boldsymbol{y}$

$$\sim G(\frac{1}{2}\{c + \sum_i \sum_j [h(\theta_{ij}) - \boldsymbol{x}_{ij}^T \boldsymbol{\beta} - u_i]^2\}, \frac{1}{2}(d + \sum_{i=1}^{I} n_i))$$

(d) $R_u \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{u}, r, \boldsymbol{y}$

$$\sim G(\frac{1}{2}(a + \sum_{i=1}^{I} u_i^2), \frac{1}{2}(b + \sum_{i=1}^{I} n_i)),$$

(e) $\theta_{ij} \mid \boldsymbol{\beta}, \boldsymbol{u}, r_u, r, \boldsymbol{y}$

$$\overset{ind}{\sim} \pi(\theta_{ij} \mid \boldsymbol{\beta}, \boldsymbol{u}, r_u, r, \boldsymbol{y})$$

$$\propto \exp\left\{\frac{y_{ij}\theta_{ij} - \phi(\theta_{ij})}{\phi_{ij}} - \frac{r}{2}[h(\theta_{ij}) - \boldsymbol{x}_{ij}^T \boldsymbol{\beta} - u_i]^2\right\} h'(\theta_{ij})$$

$E(\theta_{ij} \mid \boldsymbol{y})$, $V(\theta_{ij} \mid \boldsymbol{y})$, $cov(\theta_{ij}, \theta_{i'j'})$ $(i, j) \neq (i', j')$등의 $\boldsymbol{\theta}$에 관한 추론은 MCMC(Monte Carlo Markov Chain) 방법을 이용할 수 있으며, 깁스 샘플링(Gibbs sampling)은 이러한 MCMC 방법의 일종이다.

위의 (a)~(e)의 조건부 분포를 이용한 깁스 알고리즘은 다음의 절차에 의해 이루어진다.

(i) $\theta_{ij} = \theta_{ij}^{(0)}$, $u_i = u_i^{(0)}$, $r = r^{(0)}$를 초기값으로 하여 위의 (a)로부터 $\boldsymbol{\beta}$를 생성하고 이것을 $\boldsymbol{\beta}^{(1)}$ 이라하자.

(ii) $\beta = \beta^{(1)}$, $\theta_{ij} = \theta_{ij}^{(0)}$, $r = r^{(0)}$, $r_u = r_u^{(0)}$ 를 이용하여 (b)로부터 $u_i = u_i^{(1)}$을 생성한다.

(iii) $\beta = \beta^{(1)}$, $u_i = u_i^{(1)}$, $\theta_{ij} = \theta_{ij}^{(0)}$를 이용하여 (c)로부터 $r = r^{(1)}$을 생성한다.

(iv) $u_i = u_i^{(1)}$을 이용하여 (d)로부터 $r_u = r_u^{(1)}$ 을 생성한다.

(v) $\beta = \beta^{(1)}$, $u_i = u_i^{(1)}$, $r = r^{(1)}$, $\theta_{ij} = \theta_{ij}^{(0)}$를 이용하여 (e)로부터 $\theta_{ij} = \theta_{ij}^{(1)}$을 생성한다($i = 1, 2, \cdots, m$).

(vi) 절차 (i)~(v)를 한 사이클로 하여 같은 과정을 반복 수행한다.

수렴이 이루어지는 시점 t까지 충분히 반복한 후, 이 후부터 얻어지는 J개의 표본 $\{ \beta^{(t+k)}, u_1^{(t+k)}, u_2^{(t+k)}, \cdots, u_I^{(t+k)}, r^{(t+k)}, r_u^{(t+k)}, \theta_{1j}^{(t+k)}, \theta_{2j}^{(t+k)}, \cdots, \theta_{Ij}^{(t+k)}; k = 1, 2, \cdots, J\}$을 $\beta, u_1, u_2, \cdots, u_I, r, r_u, \theta_{1j}, \theta_{2j}, \cdots, \theta_{Ij}$ 의 결합 사후분포로부터 얻은 표본으로 간주하며, θ_{ij} 의 사후평균, 사후분산은 위에서 생성된 J개의 표본을 이용하여 추정한다.

I개의 소지역이 주어진 경우, 각 소지역내에서 여러개의 단위가 선택되고, 선택된 단위들의 응답은 서로 독립적이며 J개의 범주로 분류될 수 있다고 할 때, p_{ijk}를 i번째 소지역에서 k번째로 선택된 단위가 j번째 범주일 확률로 정의하자 ($i = 1, 2, \cdots, I$; $j = 1, 2, \cdots, J$; $k = 1, 2, \cdots, n_i$). 이 때 i번째 소지역 내에서 k번째로 선택된 단위 $Z_{ijk}(j = 1, 2, \cdots, J)$는 J개의 범주 중 하나로 구분될 수 있고, 선택된 단위의 응답은 서로 독립인 multinomial ($\sum_j Z_{ijk}; p_{i1k}, p_{i2k}, \cdots, p_{ijk}$)분포를 따른다고 가정할 수 있다. 만약 Y_{ijk}

$(j=1,2,\cdots,J)$가 서로 독립인 $poisson(\zeta_{ijk})$ 분포를 따르고, $p_{ijk}=\frac{\zeta_{ijk}}{\sum_{j=1}^{J}\zeta_{ijk}}$ $(j=1,2,\cdots,J)$일 경우, multinomial 분포와 $poisson$분포간에는 $(Z_{i1k}, Z_{i2k}, \cdots, Z_{ijk}) \overset{d}{=} (Y_{i1k}, Y_{i2k}, \cdots, Y_{ijk}) | \sum_j Y_{ijk}=t_{ik}$ 의 관계가 성립한다.

θ_{ijk}를 $\theta_{ijk}=\log\zeta_{ijk}$로 놓았을 때 $\zeta_{ijk}=\exp(\theta_{ijk})$이므로 $p_{ijk}=\frac{\exp(\theta_{ijk})}{\sum_{j=1}^{J}\exp(\theta_{ijk})}$로 표현될 수 있다. 여기에서의 관심은 p_{ijk}에 관한 추론이며, 이에 앞서 p_{ijk}의 추론에 필요한 θ_{ijk}의 모형을 다음과 같은 모형으로 구조화할 수 있다.

$$h(\theta_{ijk}) = \boldsymbol{x}_{ijk}^{T}\boldsymbol{\beta} + u_{ij} + \varepsilon_{ijk} \tag{5.53}$$

$$(i=1,2,\cdots,I;\ j=1,2,\cdots,J;\ k=1,2,\cdots,n_i)$$

여기에서 $\boldsymbol{\beta}=(\beta_1,\beta_2,\cdots,\beta_p)^T$는 $p\times1$벡터, $\boldsymbol{x}_{ijk}=(x_{ijk1}, x_{ijk2}, \cdots, x_{ijkp})^T$는 $p\times1$ 계획벡터이다. 오차 항 ε_{ijk}는 $\varepsilon_{ijk} \sim iid\ N(0, \sigma^2)$이고, 랜덤 항 u_{ij}는 $u_{ij} \sim iid\ N(0, \sigma_u^{\,2})$이며, ε_{ijk}와 u_{ij}는 서로 독립이라 가정한다.

식(5.51)과 (5.53)을 계층적인 구조의 모형으로 다음과 같이 구조화한다. 다음의 표현에서 $\boldsymbol{\theta}=(\theta_{111}, \theta_{112}, \cdots, \theta_{11n_1}, \cdots, \theta_{IJ1}, \theta_{IJ2}, \cdots, \theta_{IJn_I})^T$, $R=\sigma^{-2}$, $R_u=\sigma_u^{\,-2}$, $\boldsymbol{u}=(u_{11}, u_{12}, \cdots, u_{1J}, \cdots, u_{m1}, u_{m2}, \cdots, u_{mJ})^T$를 나타낸다.

(I) $Y_{ijk}|\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{u}, R_u=r_u, R=r$

$$\overset{ind}{\sim}\ f(y_{ijk}|\theta_{ijk}, \phi_{ijk}) = \exp\left\{\frac{y_{ijk}\theta_{ijk}-\phi(\theta_{ijk})}{\phi_{ijk}} + \rho(y_{ijk};\phi_{ijk})\right\}$$

(Ⅱ) $h(\theta_{ijk}) \mid \boldsymbol{\beta}, \boldsymbol{u}, R_u = r_u, R = r$

$$\overset{\text{ind}}{\sim} N(\boldsymbol{x}_{ijk}^T \boldsymbol{\beta} + u_{ij}, r^{-1})$$

(Ⅲ) $u_{ij} \mid \boldsymbol{\beta}, R_u = r_u, R = r$

$$\overset{\text{ind}}{\sim} N(0, r_u^{-1})$$

여기에서 모수 $\boldsymbol{\beta}, R_u = r_u, R = r$ 의 사전분포를 다음과 같이 가정한다

(Ⅳ) $\boldsymbol{\beta} \sim uniform(R^p), (p < m)$,

$$R_u \overset{\text{ind}}{\sim} G(\frac{a}{2}, \frac{b}{2}),$$

$R \sim$ gamma $(\frac{c}{2}, \frac{d}{2})$ 이고, $\boldsymbol{\beta}$, R_u, R은 서로 독립

$p_{ijk} = \dfrac{\exp(\theta_{ijk})}{\sum_{j=1}^{J} \exp(\theta_{ijk})}$ 의 사후평균, 분산, 공분산을 계산하기 위해서 필요한 사후분포들은 다음과 같다.

(a) $\boldsymbol{\beta} \mid \boldsymbol{\theta}, \boldsymbol{u}, r_u, r, \boldsymbol{y}$

$$\sim N_p\left(\frac{\sum_{i,j,k}[h(\theta_{ijk})\ \boldsymbol{x}_{ijk} - u_{ij}]}{\sum_{i,j,k} \boldsymbol{x}_{ijk}\ \boldsymbol{x}_{ijk}^T}, \ r^{-1}\left(\sum_{i,j,k} \boldsymbol{x}_{ijk}\ \boldsymbol{x}_{ijk}^T\right)^{-1}\right)$$

(b) $\boldsymbol{u}_{ij} \mid \boldsymbol{\theta}, \boldsymbol{\beta}, r_u, r, \boldsymbol{y}$

$$\overset{ind}{\sim} N\left((r n_i + r_u)^{-1} r \sum_j (h(\theta_{ijk}) - \boldsymbol{x}_{ijk}^T \boldsymbol{\beta}) , (r n_i + r_u)^{-1} \right)$$

(c) $R \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{u}, r_u, \boldsymbol{y}$

$$\sim \text{gamma}\left(\frac{1}{2}\left\{c + \sum_{i,j,k} [h(\theta_{ijk}) - \boldsymbol{x}_{ijk}^T \boldsymbol{\beta} - u_{ij}]^2\right\}, \frac{1}{2}(d + J\sum_i n_i)\right)$$

(d) $R_u \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{u}, r, \boldsymbol{y}$

$$\sim G\left(\frac{1}{2} (a + \sum_i \sum_j u_{ij}^2) , \frac{1}{2} (b + IJ) \right) , \quad s = 1, 2, \cdots, p$$

(e) $\theta_{ijk} \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{u}, r_u, r, \boldsymbol{y}$

$$\overset{ind}{\sim} \pi(\theta_{ijk} \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{u}, r_u, r, \boldsymbol{y})$$

$$\propto \exp\left\{ \frac{y_{ijk}\theta_{ijk} - \phi(\theta_{ijk})}{\phi_{ijk}} - \frac{r}{2}[h(\theta_{ijk}) - \boldsymbol{x}_{ijk}^T \boldsymbol{\beta} - u_{ij}]^2 \right\} h'(\theta_{ijk})$$

5.4.3 추정 결과

모형 (5.53)에서 k를 연령과 성별을 나타내는 (a, s)로 구분하여 표기하도록 하고, 회귀방정식 $\boldsymbol{x}_{ijk}\boldsymbol{\beta} = \mu + \tau_a^A + \tau_s^S + \tau_j^J + \tau_{as}^{AS} + \tau_{aj}^{AJ} + \tau_{sj}^{SJ}$ 을 적합시켜 소지역에서의 특성별 추정값들을 계산하고자 한다. 여기에서 μ는 일반효과, τ_a^A는 a번째 연령그룹에서의 주효과, τ_s^S는 s번째 성별그룹에서의 주효과, τ_j^J는 j번째 응답범주와 관련된 주효과, τ_{as}^{AS}는 a번째 연령그룹과 s번째 성별그룹의 교호작용의 효과, τ_{aj}^{AJ}는 a번째 연령그룹과 j번째 응답범주의 교호작용의 효과, τ_{sj}^{SJ}는 s번째 성별그룹과 j번째 응답범주의 교호작용의 효과를 나타낸다. 모든 a, s, j에 대

하여 $\tau_1^A = \tau_1^S = \tau_1^J = \tau_{a1}^{AS} = \tau_{1s}^{AS} = \tau_{a1}^{AJ} = \tau_{1j}^{AJ} = \tau_{s1}^{SJ} = \tau_{1j}^{SJ} = 0$ 의 제한조건을 가정한다.

가정된 모형을 이용한 소지역 특성별 추정결과는 다음의 <표8>과 같다. 결과적으로 경험적 베이즈 추정량의 표준오차가 표본 추정량의 표준오차에 비해 상대적으로 매우 작은 사실을 확인할 수 있다.

<표8> Impact of Exposure to Health Hazards in the Workplace

		Sample		H. Bayes	
Category	Response	Proportions	SD	Proportions	SD
Region=2	Total n=294				
M, Age<40	Yes	.400	.100	.373	.042
	No	.383	.101	.345	.041
	Not exposed	.150	.119	.199	.031
	NA/NS	.067	.125	.083	.015
F, Age<40	Yes	.257	.100	.266	.035
	No	.284	.098	.279	.035
	Not exposed	.311	.097	.274	.036
	NA/NS	.148	.107	.181	.026
M, Age≥40	Yes	.111	.111	.184	.028
	No	.153	.109	.176	.027
	Not exposed	.167	.108	.156	.026
	NA/NS	.569	.077	.484	.040
F, Age≥40	Yes	.159	.098	.110	.019
	No	.091	.102	.103	.018
	Not exposed	.125	.010	.134	.022
	NA/NS	.625	.065	.654	.034

Category	Response	Sample		H. Bayes	
		Proportions	SD	Proportions	SD
Region=3	Total n=740				
M, Age<40	Yes	.294	.070	.311	.029
	No	.426	.063	.395	.032
	Not exposed	.203	.075	.186	.023
	NA/NS	.077	.080	.108	.015
F, Age<40	Yes	.246	.064	.235	.024
	No	.273	.063	.287	.026
	Not exposed	.180	.067	.204	.023
	NA/NS	.301	.062	.274	.026
M, Age≥40	Yes	.156	.069	.154	.019
	No	.150	.069	.165	.020
	Not exposed	.100	.071	.112	.016
	NA/NS	.594	.048	.569	.028
F, Age≥40	Yes	.064	.063	.071	.010
	No	.086	.063	.091	.012
	Not exposed	.111	.062	.099	.013
	NA/NS	.739	.033	.739	.021

Category	Response	Sample		H. Bayes	
		Proportions	SD	Proportions	SD
Region=8	Total n=1707				
M, Age<40	Yes	.274	.047	.279	.021
	No	.360	.044	.362	.023
	Not exposed	.253	.048	.253	.020
	NA/NS	.113	.052	.106	.012
F, Age<40	Yes	.199	.042	.196	.016
	No	.267	.040	.275	.019
	Not exposed	.289	.040	.297	.019
	NA/NS	.245	.041	.234	.017
M, Age≥40	Yes	.113	.047	.130	.013
	No	.166	.046	.174	.016
	Not exposed	.217	.044	.195	.017
	NA/NS	.504	.035	.501	.022
F, Age≥40	Yes	.087	.042	.076	.009
	No	.123	.041	.110	.011
	Not exposed	.119	.041	.131	.012
	NA/NS	.671	.025	.683	.017

6. 기타 소지역 추정 연구

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

소지역 추정 문제에 시계열 모형이 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 가정하며, 연결모형은 다음과 같이 표현된다.

$$\theta_{it} = \boldsymbol{x}_{it}^T \boldsymbol{\beta} + v_i + u_{it} \tag{6.1}$$

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. ε_{it}는 v_i와는 독립이며 평균이 0, 분산이 σ^2 인 정규분포를 따른다. 식 (6.1)과 같은 연결 모형이 계량경제학 분야에서 폭넓게 연구되고 있다.

You and Rao(1999)는 서로 다른 세가지의 2 수준 모형((1)동일 오차분산을 갖는 모형, (2)랜덤 오차 분산을 갖는 모형, (3) 동일하지 않은 오차 분산을 갖는 모형)들에 대해 계층적 베이즈 방법을 적용하였을 때, (1), (2) 보다는 (3)이 훨씬 자료를 잘 적합시킨다는 결과를 보여 주었다. Datta, Day and Basawa(1999)는 기초적인 단위 수준(unit-level) 모형을 다변량의 경우로 확장하여 다변량 내포오차회귀모형을 이끌어 냈다.

최근에는 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 $\frac{\theta_{ij}}{1-\theta_{ij}} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + v_i$ 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.

7. 미국의 소지역 실업통계 작성

미국의 노동력 조사를 위한 표본관리체계는 4-8-4 system으로서 표본조사 가구로 한번 선정하면 4개월간 계속 조사를 실시하고 8개월간은 조사를 중지한 후, 다시 표본가구로 편입하여 4개월간 연속조사를 실시하는 제도이다. 이는 월간 노동력 변화와 연간 노동력 변화추세에 대한 추정값의 신뢰성을 높이고 표본조사 가구의 응답 부담을 줄이기 위한 제도이다. 먼저 주 단위의 실업통계 작성의 발전과정을 알아보고, 다음에는 주의 실업통계 작성절차에 대해서 설명한다.

7.1 실업통계 발전과정

1950년대에 노동성의 고용 훈련 행정국에서 각 주 간에 실업률의 추정값을 비교할 수 있도록 실업자의 추정기법을 개발하여 책자로 발간하였다. 이 책자(handbook)를 근거로 하여 50년대 후반에는 대규모의 통계조사를 위한 비용 부담을 없애고, 소규모 조사를 통해서 연속적인 여러 단계를 거치는 주 단위 실업통계 작성기법을 "Handbook"방법으로 공식화하였다. 1950년대의 실업통계 작성은 주로 실업보험 신청자료를 이용하였다.

1972년에는 노동 통계국에서 주 단위의 이용 가능한 노동력, 취업과 실업의 추정에 대한 방법과 개념을 연구하기 시작하였고, 1973년에는 통계국이 주관이 되어, 경상인구조사(CPS : Current Population Survey)의 개념, 정의, 추정과 handbook방법을 결합하여 주 단위와 주의 세부단위까지의 노동력을 추정할 수 있는 기법을 개발하였다.

1976년 이후에는 모든 주 단위 실업통계의 추정값에 대한 신뢰도를 높이기 위해서 각 주별로 표본 가구 수를 몇 배씩 증가시켰으며, 이후부터 경상인구조사의 자료를 이용한 노동력의 추정값을 공식적으로 발표하기 위해 실업률을 6%라고 가정했을 때, 실업통계에 대한 변동계수의 최대 예상 허용값을 10%로 하였다. 1978년부

터 규모가 큰 10개주(California, Florida, Illinois, Massachusetts, Michigan, New Jersey, New York, Ohio, Pennsylvania, Texas)와 2개 대도시(Los Angels, New York City)의 노동력 통계는 CPS자료에서 직접 추정법으로 계산한 결과를 공식 통계로 사용토록 하였다.

실업보험의 데이터베이스를 계속적으로 개선하였으며, 특히 1976~78년 사이에 실업보험 신청자료를 모든 주에 대해서 표준화하고, CPS조사의 조사주간(매월 12일을 포함한 주)에 실업자로 인정받는 실업 보험 신청자는 자동으로 데이터베이스에 등록되도록 하는 데이터 관리체계를 개발하여 CPS자료와 연계한 추정법을 개발하였다.

1985년에는 1980년 센서스 정보를 이용하여 state-based CPS 표본설계를 완성하였으며, North Carolina주를 CPS 자료에서 노동력 통계를 직접 추정하여 공표 하는 주로 포함시켰고, 대규모 11개 주(노동력 통계를 직접 추정법으로 추정하는 주)의 월별 실업통계에 대한 목표변동계수를 8%로 낮추었다. 또한 나머지 39개 주의 연평균 실업통계에 대한 목표변동계수를 8%로 정하였으며, 이 때 실업률의 참값은 6%인 것으로 가정하였다.

1989년까지는 직접 추정법을 적용하지 않은 39개 주와 substate의 실업수준의 공식적 월별 추정값은 Handbook방법을 적용하여 계산되었으나, 1989년 초부터 39개 주에 대한 노동력 통계 작성에서는 노동 통계국에서 개발한 시계열 모형 기법을 표준 추정법으로 채택하였다.

1994년에는 좀 더 고급회귀모형을 개발하여 소규모의 39개 주에 대한 실업통계 작성에 적용하였다.

1992년에는 주 단위의 추정값에 계절 조정을 적용하였고 1994년에는 1990년 센서스 자료를 이용하여 CPS를 재 설계하였으며, 또한 모든 조사에서 컴퓨터 보조면접을 실시할 수 있도록 설문지를 재구성하였다. 새로 개편된 CPS의 표본설계는 1995년 중반까지 단계적으로 도입하였다.

1996년부터는 예산 절감으로 CPS표본 규모를 56,000가구에서 50,000가구로 축소했기 때문에 모든 주의 실업통계는 CPS자료에서 직접 추정하는 방식을 적용하지 않고 시계열 회귀모형을 적용하여 작성하도록 하였으며, 로스엔젤스와 뉴욕시의 실업통계 생산에 계절 조정 자료를 적용하였다. 그러나 소지역 단위인 LMA(labor market area)의 실업통계를 생산하는데 좀 더 간편화된 회귀모형 같은 통계적 모형을 적용하기보다는 해당 LMA의 취업과 실업의 구조적 분석을 통한 building block절차를 적용하지만, 마지막 주정부내의 합계는 공식적인 주단위의 실업통계와 일치하도록 비례조정을 하였다.

7.2 회귀 모형 추정법

회귀모형을 이용한 추정법은 합성 추정값이 지역 변동을 충분히 설명하지 못한 점을 보완하고자 실직 보험 신청자료와 구직 등록 신청자료를 이용하여 1970년대 후반부터 시군구의 실업통계 작성에 이용되었던 기법이다. CPS 1차 추출단위(PSU : Primary Samplint Unit)의 추정치를 종속변수로 사용하고, 다음과 같은 몇 개의 적절한 독립변수를 선택한 회귀모형을 고려한다.

(1) 다음의 범주에서의 합성추정치

① Occupation-Sex-Race

② Marital Status-Age-Sex-Race

(2) 센서스에서 추정한 실업자 총계 대비 3-4월의 실직보험 가입자에 대한 가입비율(%)

(3) 연말 자료에서 공표되는 "70-step"의 실업자 추정치(실업률)

5개월 간의 월별 CPS 추정값들의 평균을 종속변수로 사용하기 위해, CPS-PSU와 SMSA를 대응시켰을 때 150개의 SMSA 중 122개의 SMSA가 완전 대응되며 이러한 지역을 회귀모형 추정을 위한 자료로 활용하였다. 실업률 추정을 위해 적합한 회귀모형은 다음과 같다.

$$\hat{Y} = 0.008 - 0.201 X_1 + 0.680 X_2 + 0.404 X_3 \ , \qquad (7.1)$$

Residual Mean Square = 0.868×10^{-4} ,

추정치의 표준오차 = 0.932×10^{-2} ,

$R^2 = 0.546$,

여기에서 Y = '5개월 간의 월별 CPS 실업추정치의 평균', X_1 = '전체 실업자 대비 보험가입 실업자의 비율(%)', X_2 = '최종 공표된 연말 실업률(final annual 70-step estimates)', X_3 = 'Marital Status-Age-Sex-Race 범주에서 계산된 실업률의 합성추정치'를 나타낸다.

위 회귀모형의 타당성 검증을 위하여 센서스 자료를 이용한 회귀모형의 적합결과는 다음과 같이 주어지며, 실업률 추정에서 회귀모형의 적용가능성을 보여준다.

$$\hat{U} = 0.009 + 0.012 X_1 + 0.586 X_2 + 0.540 X_3 \ , \qquad (7.2)$$

Residual Mean Square = 0.213×10^{-4} ,

추정치의 표준오차 = 0.461×10^{-2} ,

$R^2 = 0.859$,

여기에서 U는 센서스 자료에서 계산한 실업률 추정치이다.

5개월 간의 월별 CPS 실업 추정치의 평균을 종속변수로 사용했을 때, 변동의 55%가 회귀모형으로 설명될 수 있고, 센서스의 실업률 추정치를 종속변수로 사용한 경우에는 변동의 86%가 회귀모형으로 설명될 수 있다. CPS 추정치의 표본변동이 커질수록 설명변수들의 기여도가 낮아진다는 것을 알 수 있다.

독립변수를 변화시켜 가면서 추가로 2종의 회귀모형을 추정해 보기로 한다. 먼저 독립변수로써 X_4를 CPS자료에 의한 추정치를 benchmarking하기 전의 실업 추정값으로 취하여 회귀모형을 적합했을 때 다음과 같은 회귀모형을 얻을 수 있다.

$$\hat{Y} = 0.012 - 0.252X_1 + 0.299X_3 + 0.078X_4 \quad , \tag{7.3}$$

Residual Mean Square = 0.833×10^{-4} ,

추정치의 표준오차 = 0.912×10^{-2} ,

$R^2 = 0.565$

대응되는 회귀모형은 다음과 같다.

$$\hat{U} = -0.006 + 0.005X_1 + 0.477X_3 + 0.564X_4 \quad , \tag{7.4}$$

Residual Mean Square = 0.228×10^{-4} ,

추정치의 표준오차 = 0.478×10^{-2} ,

$R^2 = 0.849$

식(7.3)의 회귀모형은 총변동의 약 57%를 설명하며, 회귀모형 (7.1)보다는 어느 정도 개선되었다.

두 번째 경우에는 최종 공표된 연말 실업률 X_2대신 월별 추정치의 11개월 간의 평균 X_5와 Occupation-Sex-Race의 3개의 범주에서 구한 합성추정치 X_6를 독립변수로 하였을 때 계산된 회귀모형의 적합 결과이다.

$$\hat{Y}'' = 0.009 - 0.210X_1 + 0.640X_5 + 0.444X_6 \quad , \tag{7.5}$$

Residual Mean Square = 0.883×10^{-4} ,

추정치의 표준오차 = 0.939×10^{-2} ,

$R^2 = 0.539$

센서스 자료 추정치를 이용한 적합 결과는 다음과 같다.

$$\hat{U}'' = -0.008 - 0.011X_1 + 0.532X_5 + 0.617X_6 \quad , \tag{7.6}$$

Residual Mean Square = 0.194×10^{-4} ,

추정치의 표준오차 = 0.440×10^{-2} ,

$R^2 = 0.872$

이상에서 살펴보았듯이 센서스 자료의 추정치를 종속변수로 사용했을 경우 회귀식은 약 85~87%의 설명력을 보이나, CPS자료 추정치를 사용했을 경우에는 표본추출에 의한 변동에 기인하여 회귀식의 설명력은 약 54~56%의 범위 정도에 있음을 확인할 수 있다.

<표8>은 변수들 간의 상관관계를 나타낸 표이며, 70-step 추정치들은 센서스 추정치와 CPS 추정치들과 높은 상관관계를 보임을 알 수 있다. 또한, 합성추정치와 실직보험가입비율 추정치와는 낮은 상관성을 나타내며, 이들과 70-step 추정치들과도 상관계수의 값이 낮다. 따라서 추가적인 독립변수로써 합성추정치와 실직보험가입비율을 선택한다면 회귀식의 예측력은 증가할 것임을 알 수 있다.

<표 8> 변수들간의 가중 상관계수

	X_1	X_2	X_3	X_4	X_5	X_6	Y	Z	U
X_1	1.000								
X_2	0.676	1.000							
X_3	0.285	0.512	1.000						
X_4	0.682	0.961	0.574	1.000					
X_5	0.666	0.995	0.477	0.959	1.000				
X_6	0.369	0.584	0.974	0.633	0.548	1.000			
Y	0.372	0.692	0.577	0.720	0.682	0.599	1.000		
Z	0.259	0.525	0.340	0.543	0.521	0.339	0.700	1.000	
U	0.554	0.851	0.756	0.868	0.835	0.810	0.741	0.512	1.000

* Z = 1970년 5월 CPS 실업추정치

센서스 중간 해에는 회귀식의 종속변수는 CPS 1차 추출단위의 추정치가 되며, 독립변수로 센서스에서 추정한 실업률(U)을 선택할 수 있으며, 적합된 회귀식은 다음과 같다.

$$\hat{Y} = 0.010 + 0.450\,U + 0.326\,X_4 + 0.089\,X_6 \quad , \tag{7.7}$$

Residual Mean Square = 0.835×10^{-4} ,

추정치의 표준오차 = 0.914×10^{-2} ,

$R^2 = 0.563$

만약 종속변수로 1970년 5월 CPS 실업 추정치(Z)를 사용한다면 적합된 회귀모형은 다음과 같다.

$$\hat{Z} = 0.019 + 0.422\,U + 0.430\,X_4 - 0.246\,X_6 \quad , \tag{7.8}$$

$$\text{Residual Mean Square} = 2.040 \times 10^{-4} \ ,$$

$$\text{추정치의 표준오차} = 1.428 \times 10^{-2} \ ,$$

$$R^2 = 0.291$$

모형 적합시 센서스 시기뿐만 아니라 센서스 중간 해의 소지역의 추정치에 대한 자료가 필요하다. 센서스 중간 해에 종속변수로 CPS 추정치를 이용하고, 독립변수로 센서스 추정치를 이용한 회귀모형 사용이 가능하다. 식 (7.7)은 독립변수로 센서스 추정치를 이용한 결과를 보여준다. 회귀모형은 약 56%의 설명력을 가지며, 또한 Y를 종속변수로 갖는 다른 모형과 비교했을 때 결코 떨어지지 않는 설명력을 갖는다.

1970년 5월 CPS 실업 추정치를 종속변수로 취한 회귀식 (7.8)은 약 29%의 설명력을 갖는다. 이러한 낮은 설명력은 종속변수의 큰 분산(변동)에 기인한다.

노동력의 규모가 서로 상이한 지역들에 대한 잔차분석 결과가 <표9>에 주어졌다. 독립변수가 각각 (X_1 , X_3 , X_4), (X_1 , X_5 , X_6)이고, 종속변수로써 각각 1970년 센서스 추정치, 5개월간의 월별 CPS 추정치, 1970년 5월의 CPS 추정치를 이용한 회귀모형에 대한 잔차의 평균과 잔차의 표준오차이다.

<표9> 잔차의 평균과 표준오차 (단위 : %점)

1970 Census Labor Force	Number of SMSA's	Census unemployment		CPS-five month average unemployment		CPS-April 1970 unemployment	
		Mean residual	S.E of residual	Mean residual	S.E of residual	Mean residual	S.E of residual
(X_1, X_3, X_4)							
1,000,000 이상	9	-0.18	0.31	-0.01	0.39	-0.00	0.54
500,000-999,999	17	-0.02	0.54	-0.06	0.66	0.24	0.88
250,000-499,999	21	0.15	0.36	0.15	1.17	-0.21	1.23
100,000-249,999	51	0.14	0.66	0.03	1.55	0.10	2.77
100,000 미만	24	0.37	0.78	0.41	1.83	-0.46	2.83
(X_1, X_5, X_6)							
1,000,000 이상	9	-0.07	0.34	0.07	0.38	0.06	0.55
500,000-999,999	17	-0.10	0.49	-0.16	0.75	0.10	0.90
250,000-499,999	21	0.06	0.39	0.06	1.23	-0.28	1.37
100,000-249,999	51	0.10	0.61	0.03	1.57	0.14	2.80
100,000 미만	24	0.11	0.71	0.19	1.75	-0.63	2.86

노동력의 규모가 작은 지역들보다는 노동력의 규모가 큰 지역들에서 추정치의 효율이 좋게 나타난다. 1970년 Cessus 추정치를 이용한 회귀모형을 제외하면, 잔차의 표준오차는 노동력의 규모가 작을수록 증가하는 경향을 보인다. Census 추정치를 종속변수로 취한 회귀모형에서 노동력의 규모가 100,000이하인 지역에서의 잔차의 표준오차는 노동력의 규모가 백만 이상인 지역보다 2배 이상 큰 표준오차를 갖는다. 5개월 간의 월별 CPS 평균에 대한 추정치를 이용한 회귀모형에서의 잔차의 표준오차를 검토해 볼 때, 1970년 5월 CPS 추정치를 이용한 회귀모형 보다는 잔차의 표준오차가 훨씬 작음을 알 수 있다.

1970년 센서스 실업률과 1970년 센서스 실업 추정치를 이용한 회귀모형 추정치 간의 차이에 대한 분포가 <표10>에 주어졌다. 독립변수는 (X_1 , X_5 , X_6)를 이용하였다. <표8>을 살펴보면 센서스 실업 추정치를 이용했을 경우 CPS 추정치를 이용한 경우보다 차이의 분포가 훨씬 대칭적임을 알 수 있다.

−1.0~1.0 %점의 구간을 살펴보면, 센서스 회귀모형의 경우는 91.9%, CPS-5-month의 경우는 75.4%, CPS-April의 경우는 73.0%가 위치함을 발견할 수 있으며, −0.50%점 아래 구간에서는 CPS 회귀식이 큰 비율로 분포해 있음을 알 수 있다(CPS-5-month의 경우 62.2%, CPS-April의 경우 59.8%가 위치함).

<표10> 센서스실업률과 회귀추정치간의 차이에 대한 분포(122개의 SMSA 이용)

Difference classes (% point)	Census unemployment		CPS-five-month average unemployment		CPS-April 1970 unemployment	
	No. of SMSA's	Percent of SMSA's	No. of SMSA's	Percent of SMSA's	No. of SMSA's	Percent of SMSA's
3.00 이상	0	0	0	0	0	0
2.00~3.00	1	0.8	1	0.8	1	0.8
1.50~2.00	0	0	0	0	0	0
1.00~1.50	6	4.9	0	0	2	1.6
0.50~1.00	14	11.5	4	3.3	5	4.1
0.25~0.50	15	12.3	5	4.1	8	6.6
0.10~0.25	19	15.6	1	0.8	3	2.5
-0.10~0.10	23	18.9	9	7.4	6	4.9
-0.25~-0.10	10	8.2	11	9.0	8	6.6
-0.50~-0.25	15	12.3	15	12.3	16	13.1
-1.00~-0.50	16	13.1	47	38.5	43	35.2
-1.50~-1.00	3	2.5	25	20.5	23	18.9
-2.00~-1.50	0	0	2	1.6	5	4.1
-3.00~-2.00	0	0	2	1.6	2	1.6
-3.00 이하	0	0	0	0	0	0

5개월 간의 월별 CPS 추정치 평균을 종속변수로 취한 회귀모형에 대한 MSE의 추정치 $MSE = E\left\{\frac{(Y_0 - \hat{Y})^t (Y_0 - \hat{Y})}{n}\right\} - \frac{(n-2p-2)\,\sigma_w^2}{n}$ 을 이용하여 계산된 결과가 다음에 주어졌다. 단, Y_0 = '관측값 Y', n = '관측값의 개수', p = '독립변수의 수', σ_w^2 = 'PSU 내의 오차'를 나타낸다.

Independent Variables	*MSE*
X_1 X_2 X_3	0.405×10^{-4}
X_1 X_3 X_4	0.369×10^{-4}
X_1 X_5 X_6	0.419×10^{-4}

그러나 이러한 MSE의 수치들은 단순한 대략적인 추정결과이다. 왜냐하면 독립변수들 간의 상관계수가 상당히 높은 값을 갖고 있기 때문이다.

상기한 자료들은 다음과 같은 내용을 시사한다. 미 노동부에서 연말 공표하는 실업 추정치는 CPS의 1차 추출단위 추정치를 독립변수로 취한 회귀모형을 이용하여 개선될 수 있음을 시사한다. 추가적인 독립변수들로 70-step 추정치 외에 실직보험 가입자료와 센서스와 CPS 자료를 이용한 합성추정치 및 센서스 중간년도의 표본 자료의 추정치 등을 이용할 수 있다.

회귀모형에 의한 소지역 실업 통계의 추정은 적절한 독립변수와 종속변수의 선택이 중요한 과제이지만, 한편, 센서스와 CPS자료를 이용한 실질적인 분석 및 실업에 관련된 사회경제적 요인파악도 중요한 과제이다.

7.3 시계열-회귀모형

7.3.1 서 론

시계열-회귀모형은 현재 미국에서 주 단위 등의 소지역 실업통계를 작성하는데 적용하고 있는 방법으로 모집단의 값을 확률과정으로 생각할 뿐만 아니라 직접 표본조사 추정량을 개선하기 위해서 시계열 분석의 signal 적출 기법을 적용한 것이다.

1989년 1월 노동통계국(BLS)에서 미국의 39개 주와 콜롬비아의 1개 구역 총 40개 주에 대해 월별 고용과 실업 추정을 위한 새로운 방법을 소개하였다. CPS의 월별 표본자료에 시계열 모형을 적합시키는 방법이다.

모집단의 특성을 추정하기 위한 직접적인 방법은 표본 설계에 근거하여 대규모 표본조사를 시행하는 것이다. 추정치들은 대 지역에 대해서는 신뢰할 만 하나 소지역에 대해서는 그러하지 못한 실정이다. 주기적인 조사의 경우 소지역에 대해서 시계열 기법이 관심을 받고 있으며, CPS 추정치는 이러한 기법을 적용시키기에 적합한 자료이다. 매달 59,000가구가 조사되어 모집단의 노동력 상태를 추정하는데, 전체 추정치 또는 인구수가 많은 11개 주에 대한 월별 추정치는 비교적 신뢰할 만하다. 인구수가 적은 나머지 40개 주에 대해서는 월별 추정치를 그대로 사용하는 것은 바람직하지 못하다. 1989년 이전에는 40개 주에 대한 노동력 추정치는 BLS 핸드북에서 제시한 방법에 의해서 수행되었다.

새로운 추정방법은 CPS 표본 자료를 확률적으로 변화하는 노동력 시계열인 signal 성분과 noise 성분의 합으로 표현되는 모형 추정에 근거한다. 표본 설계 정보에 따라 월별 CPS 노동력 추정치는 시계열 모형에서 실업보험자료(UI :Unemployment Insurance Data)와 경상고용통계자료(CES: Current Employment Statistics Data)를 결합하여 계산한다. 즉, 이전에 추정한 방법과는 달리 좀더 체계적인 방법으로써 보조자료와 과거와 현재의 표본자료를 모두 이용하여 표본크기가 작은데에서 발생하는 CPS 추정치의 큰 분산값을 줄이고자 하는 것이 기본적인

생각이다.

비 관측된 모집단 값의 동적인 변화와 표본오차의 자기공분산을 나타내는 모형이 주어진다면, Kalman Filter는 참값을 추정하기 위해 사용될 수 있으며, 다음과 같은 유용성을 갖고 있다. 첫째, KF는 signal성분과 noise성분으로 표현된 모형들에 대한 다양한 접근방법들을 허용한다. 둘째, KF의 반복적인 방법은 소지역의 월별 노동력 추정치를 산출하는데에 있어서 매우 효율이 좋은 알고리즘을 제공한다. 셋째, KF는 동적인 모형들에 있어서 미지인 모수들을 추정하기 위한 매우 유용한 도구이다.

7.3.2 CPS 자료를 모형화하기 위한 시계열 방법

CPS 노동력 추정치 $y(t)$는 signal $\theta(t)$와 noise $e(t)$의 합 $y(t) = \theta(t) + e(t)$으로 나타낸다. signal 성분(노동력의 참값)은 시간에 따라 변화하는 평균 $\mu_X(t)$와 오차항 $u(t)$로 표현할 수 있다.

$$\theta(t) = \mu_X(t) + u(t) \tag{7.9}$$

$$= X(t)\beta(t) + u(t)$$

단, $X(t) = 1\times k$ 벡터(관측된 값), $\beta(t) = k\times 1$ 벡터(확률계수 벡터) 이다.

회귀계수 $\beta(t)$는 1차 자기상관회귀 과정에 따라 확률적으로 변화하는 형태를 수식으로 표현하면 아래와 같다.

$$\beta(t) = T_\beta \beta(t-1) + v_\beta(t) \tag{7.10}$$

여기에서 $T_\beta = k\times k$ 행렬(고정모수들), $v_\beta(t) = k\times 1$ 벡터(백색잡음오차)이다.

식(7.9)에서 오차 $u(t)$는 자기상관성을 고려하여 $u(t) \sim ARMA(p_u, q_u)$를 가정하여 다음과 같이 표현할 수 있다.

$$u(t) = \phi_u(L)\ \ \Psi_u(L)^{-1}\, v_u(t) \tag{7.11}$$

여기에서 $u(t)$ = 방정식의 오차, $v_u(t) = u(t)$에 대한 백색잡음오차,

$$\Psi_u(L) = 1 - \sum_{i=1}^{p_u} \Psi_u(i)\, L^i\ , \quad (\ u(t)\text{에 대한 자기회귀연산자})$$

$$\phi_u(L) = 1 + \sum_{i=1}^{q_u} \phi_u(i)\, L^i\ , \quad (\ u(t)\text{에 대한 } MA \text{ 연산자})$$

L = 시차연산자(lag operator)로써 $L^i y(t) = y(t-i)$를 만족하고,

랜덤 오차 $v_\beta(t)$와 $v_u(t)$는 평균이 0이고 서로 독립을 가정한다. 즉,

$$\begin{bmatrix} v_\beta(t) \\ v_u(t) \end{bmatrix} \sim \mathrm{ID}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} Q & 0 \\ 0 & \sigma_{v_u v_u} \end{bmatrix} \right) \tag{7.12}$$

단, $Q = Cov(v_\beta(t)) = Diag(\sigma_{\beta_1\beta_1}, \cdots, \sigma_{\beta_k\beta_k})$

noise 성분은 전체 모집단에서 일부를 표본 추출하는 과정에서 발생하는 오차인 표본추출오차로 취급된다. CPS는 모집단으로부터 추출된 복잡한 다 단계 표본이다. 첫 단계에서 1차 추출단위(PSU's)들의 층화된 표본이 뽑히고, 다음으로 PSU 내에서 가구단위들이 층화된 집락 표본으로부터 뽑히며, 이때 매달 가구조사에서 일부 가구들은 대체 가구로 바뀌어 조사된다. CPS 추정과정은 비면접 보정인 2단계의 비율 보정과 중복표본을 고려한 합성절차로 이루어져 있다(Bureau of the Census(1978)).

$y(t)$가 노동력 특성을 나타내는 $\theta(t)$에 대한 CPS 추정치로 주어지면, 표본오차 $e(t) = y(t) - \theta(t)$의 분산과 공분산 함수는 다음과 같다.

$$\sigma_{e(t)e(t)} = D_y\ S_y^2\ , \tag{7.13}$$

$$\gamma_{ts} = Cov(e(t)\ ,\ e(s))\ ,$$

여기에서 D_y = 단순임의추출표본(SRS) 추정치의 분산에 대한 CPS 추정치 분산의 비(Design Effect),

$$S_y^2 = \frac{N(t)}{n(t)}\ \theta(t)\,(1-P(t)),$$

$N(t)$ = 모집단 크기,

$n(t)$ = 표본 크기,

$$P(t) = \frac{\theta(t)}{N(t)}$$

위의 식에서 설명한 것과 같이 CPS 표본오차는 이분산적 구조와 자기상관적 구조를 동시에 갖는다. 식(7.13)은 다음과 같은 이분산성의 세가지 주요한 원인을 나타낸 식이다. 즉, 표본 재설계에서 설계효과 D_y, 표본 구간 $\frac{N(t)}{n(t)}$, 참값 $\theta(t)$와 $P(t)$의 변화가 반영되었다.

부연 설명하면, CPS는 10년 간의 센서스 자료를 사용하여 매 십년 마다 표본 추출 구조 및 추정 절차를 갱신하기 위하여 재 설계된다. 1984-1985년 동안 비율보정과 합성과 같은 비 면접에 의한 개선된 추정 절차가 주 단위의 설계로 시행되었다. 표본조사의 재 설계보다는 오히려 주 단위의 표본 크기의 조정이 자주 있었고, 이것은 주 수준의 분산에 중요한 영향을 끼치게 되었다. 일종의 고정된 설계와 고정된 표본 크기일지라도 오차 분산은 참 노동력 크기의 함수이므로 변하게 된다. 노동력은 아주 순환적이고 동시에 계절적이므로 분산도 이와 비슷한 형태를 가진다는 것을 예측할 수 있다.

$e(t)$의 자기공분산 구조는 다음의 3가지 사실에 기인한다. 첫째, 월별 표본은 8개의 독립인 연동교체그룹들인 부 표본들로 구성되어 있다. 연동교체그룹들은 4개월 간 연속 조사되고, 8개월 동안 조사되지 않다가 다시 4개월 간 연속 조사 후 표본에서 영구히 제거된다. 이 과정에서 동일 주택단위들이 나타나므로 당연히 상관

성이 존재한다. 4-8-4 연동교체 체계는 월별 추정값의 신뢰성을 높이기 위해 15개월 주기 동안 첫 달은 전월에 조사된 표본의 75%를 중복하여 월 표본으로 사용하고, 나머지는 교체표본 추출방식에 의해 새로운 표본을 조사한다. 2월과 12월을 살펴보면 중복율을 50%이다. 둘째, 연동교체 체계의 사용은 표본들에 대한 주기적 선택을 요구한다. 한 집락의 주택단위들이 연동교체그룹으로부터 영구히 제외될 때, 근처에 있는 단위들로 대체된다. 따라서 새로운 단위들도 대체될 단위들과 비슷한 특성을 가질 것이므로 같은 연동교체그룹에서 다른 가구들과도 높은 상관성을 갖게된다. 셋째, 표본 오차의 변화성은 복합추정량에 의해 영향을 받게된다.

CPS 추정량의 이분산적이며 자기상관적 구조는 $e(t)$를 $e(t) = \gamma(t)\, e^*(t)$와 같은 승법적 구조로 모형화 함으로써 설명될 수 있다(Bell and Hillmer(1989)). 여기에서 $e^*(t)$는 $ARMA$과정을 따르며 상수인 분산을 갖는다.

$$e^*(t) = \phi_e(L)\ \Psi_e(L)^{-1}\ v_e(t) \tag{7.14}$$

여기에서 $v_e(t) \sim \text{NID}\ (0\ ,\ \sigma_{v_e v_e})$, $\sigma_{e^*(t)\,e^*(t)} = \sigma_{v_e v_e} \sum_{k=0}^{\infty} g_k^{\,2}$, 가중치 $\{g_k\}$는 생성함수 $g(L) = \phi_e(L)\ \Psi_e(L)^{-1}$로 계산된다. $e(t)$의 이분산성 성분 $\gamma(t) = \sqrt{\dfrac{\sigma_{e(t)\,e(t)}}{\sigma_{e^*(t)\,e^*(t)}}}$ 는 분산비의 제곱근이다. $e(t)$의 자기상관구조는 표본 설계에 의해 영향을 받게된다. 예를 들면, 표본 재 설계에서 복합추정량에서의 가중치들이 모두 교체되는 것과 같은 경우이다.

7.3.3 상태공간형식(State Space Form)과 KF 알고리즘

노동력 모형의 signal 성분과 noise 성분을 상태공간 형식에 삽입한다. 먼저 추정에는 적절치 못할지라도 융통성을 보여 주기에는 유용할 것 같은 매우 일반적인 형식을 설명하고 실질적인 적용에 있어서는 부과되어야 할 것 같은 일종의 제한들에 대해서 토의한다.

상태공간 작성에 있어서 비 관측된 signal과 noise가 상태 변수들이고, 이 들의 시간에 따른 전개는 변이함수들의 집합에 의해 설명된다. 관측짐 함수는 상태 변수들을 관측 표본 시계열들로 변환시킨다. 상태공간 체계에서 변이함수들은 1계의 VAR 형태를 취한다.

우리의 문제에서 관측할 수 없는 변수는 $\beta(t)$, $u(t)$, $e^*(t)$이다. 계수벡터인 $\beta(t)$는 (7.10)식과 같이 이미 적절한 형식이 있다. $u(t)$와 $e^*(t)$는 (7.11)식과 (7.14)식에서와 같이 $ARMA$과정으로 명시되나, 각각 1계 VAR형태인 $S_u(t)$와 $S_e(t)$벡터들로 변환된다. 임의의 $ARMA(p, q)$과정은 일종의 $r\times 1$ 1계 VAR형태로 변환될 수 있다는 것이 기본적인 규칙이며, 여기에서 $r = \max(p, q+1)$이다(Harvey(1981)). 변이함수들은 다음과 같이 주어지며, 여기에서 $S(t)$는 $\beta(t)$와 $S_u(t)$, $S_e(t)$로 구성되는 상태벡터이다.

$$S(t) = T(t)S(t-1) + \Gamma(t)v(t) \quad ,$$

$$(m\times 1)\ (m\times m) \qquad (m\times l)(l\times 1)$$

$$\begin{bmatrix} \beta(t) \\ S_u(t) \\ S_e(t) \end{bmatrix} = \begin{bmatrix} T_\beta & 0 & 0 \\ 0 & T_u & 0 \\ 0 & 0 & T_e(t) \end{bmatrix} \begin{bmatrix} \beta(t-1) \\ S_u(t-1) \\ S_e(t-1) \end{bmatrix} + \begin{bmatrix} I_k & 0 & 0 \\ 0 & \Gamma_u & 0 \\ 0 & 0 & \Gamma_e(t) \end{bmatrix} \begin{bmatrix} v_\beta(t) \\ v_u(t) \\ v_e(t) \end{bmatrix} ,$$

$$E(v(t)\,v(t)^t) = block\ \ diagonal(Q, \sigma_{v_u v_u}, \sigma_{v_{e(t)} v_{e(t)}})$$

여기에서 $T_u = \begin{bmatrix} \Psi_u(1) & : & I_{r_u-1} \\ \Psi_u(2) & : & \\ \cdot & : & \\ \cdot & : & \\ \cdot & : & \\ \cdots\cdots & \cdots & \cdots\cdots \\ \Psi_u(r_u) & : & 0 \end{bmatrix}$, $T_e(t) = \begin{bmatrix} \Psi_e(1) & : & I_{r_e-1} \\ \Psi_e(2) & : & \\ \cdot & : & \\ \cdot & : & \\ \cdot & : & \\ \cdots\cdots & \cdots & \cdots\cdots \\ \Psi_e(r_e) & : & 0 \end{bmatrix}$,

$$\Gamma_u = \begin{bmatrix} 1 \\ \phi_u(1) \\ \phi_u(2) \\ \cdot \\ \cdot \\ \cdot \\ \phi_u(r_u-1) \end{bmatrix}, \quad \Gamma_e(t) = \begin{bmatrix} 1 \\ \phi_e(1) \\ \phi_e(2) \\ \cdot \\ \cdot \\ \cdot \\ \phi_e(r_e-1) \end{bmatrix},$$

$m = k + r_u + r_e$,

$k =$ 회귀변수의 수,

$l = k + 2$,

$\gamma_u = \max(p_u, q_u+1)$, p_u와 q_u는 $u(t)$의 $ARMA$ 모수,

$\gamma_e = \max(p_e, q_e+1)$, p_e와 q_e는 $e(t)$의 $ARMA$ 모수.

관측점 함수는 벡터 $H(t)$를 선택하여, signal성분과 noise성분을 만들기 위한 상태변수들의 일차결합을 취한다.

$$Y(t) = H(t)S(t) = \theta(t) + e(t) \ , \tag{7.15}$$

여기에서 $H(t) = [X(t) | 1 | 0_{r_u-1} | \gamma(t) | 0_{m-k-r_u-2}]$,

$\theta(t) = H_\theta(t)S(t)$, $e(t) = H_e(t)S(t)$,

$H_\theta(t) = [X(t) | 1 | 0_{m-k-1}]$, $H_e(t) = [0_{k+r_u} | \gamma(t) | 0_{r_e-1}]$.

비 관측된 signal과 noise성분들의 상태공간형태가 주어진다면, KF는 signal과 noise를 추정하기 위한 방법을 제공한다. 이러한 알고리즘을 설명하기 위해 시간 $t-j$까지 관측된 자료에 대한 $S(t)$의 조건부 평균 및 공분산 행렬을 다음과 같이 표현하자.

$$S(t|t-j) = E(S(t)\,|\,Y_{t-j}, \cdots, Y_1),$$

$$P(t|t-j) = E\{(S(t)-S(t|t-j))(S(t)-S(t|t-j))^t\,|\,Y_{t-j}, \cdots, Y_1)\}$$

또한 바로 전까지의 값들이 주어진 경우의 표본 추정치 $Y(t)$의 예측값을 $Y(t|t-1) = H(t)\,S(t|t-1)$, $Y(t)$ 의 분산을 아래와 같이 표현하자.

$$E(Y(t)-Y(t|t-1))^2 = H(t)\,P(t|t-1)\,H(t)^t = f(t|t-1).$$

t 번째 관측점까지의 (t번째 관측점은 제외) 자료에 근거한 $S(t)$의 추정치가 주어진다면, 최근의 자료를 고려한 $S(t)$의 추정량은 $S(t|t-1)$과 최근의 표본 추정치 $Y(t)$의 가중평균으로 다음과 같이 표현되며,

$$S(t|t) = (I-K(t)\,H(t))\,S(t|t-1) + K(t)\,Y(t)$$

$$= S(t|t-1) + K(t)\,(Y(t) - Y(t|t-1))$$

공분산 행렬은

$$P(t|t) = (I - K(t)\,H(t)^t)\,P(t|t-1)$$ 로써,

$S(t|t-1) = TS(t-1|t-1)$와 $P(t|t-1) = TP(t-1|t-1)\,T^t + \Gamma E(v(t)\,v(t)^t)\Gamma^t$ 로부터 반복적으로 추정된다. 여기에서 가중벡터 $K(t)$(gain of filter)는 $K(t) = \dfrac{P(t|t-1)\,H(t)^t}{f(t|t-1)}$ 로 표현되며, $K(t)$의 원소들은 $P(t|t)$의 대각원소들의 합을 최소화하여 결정한다(Gelb(1974)).

KF 방정식을 이용하여, 시간 t에서 관측된 표본 추정치는 signal성분과 noise성분으로 분해된다.

$$Y(t) = \theta(t|t) + e(t|t)\ ,$$

여기에서 $\theta(t|t) = \theta(t|t-1) + h_\theta(t)\ \hat{Y}(t)$,

$$e(t|t) = e(t|t-1) + (1-h_\theta(t))\,\hat{Y}(t),$$

$$\hat{Y}(t) = Y(t) - Y(t|t-1),$$

$$h_\theta(t) = H_\theta(t)\,K(t) = \frac{\left\{ Var(\frac{\theta(t)}{\theta(t-1)}) + H_\theta(t)\,TP(t-1|t-1)\,T^t H(t) \right\}}{f(t|t-1)},$$

$$1-h_\theta(t) = H_e(t)\,K(t) = \frac{\left\{ \sigma_{e(t)e(t)} + H_e(t)\,TP(t-1|t-1)\,T^t H(t) \right\}}{f(t|t-1)},$$

$$Var(\frac{\theta(t)}{\theta(t|t-1)}) = \sum_{i=1}^{k} X_i(t)^2\,\sigma_{\beta_i\beta_i} + \sigma_{v_u v_u},$$

$$\sigma_{e(t)e(t)} = \gamma(t)^2\,\sigma_{v_e v_e}.$$

가중치 $h_\theta(t)$는 예측오차 $\hat{Y}(t)$를 signal과 noise성분으로 분해하며, 이것은 KF가 signal성분의 최소평균제곱오차 추정치를 만들기 위하여 시계열 추정량 $\theta(t|t-1)$와 최근의 표본 추정치 $Y(t)$를 결합하는 방법을 설명한다.

$\theta(t|t-1)$의 $Y(t)$에 대해서 보정되는 양은 이분산성의 표본오차 분산 $\sigma_{e(t)e(t)}$와 상대적으로 비교하여 시계열 분산성분 $Var(\frac{\theta(t)}{\theta(t-1)})$의 크기의 함수로 표현된다. $\sigma_{e(t)e(t)}$의 값이 크면 $h_\theta(t)$의 값이 작게되며, 따라서 $\theta(t|t)$를 이끌어 내는데 있어서 시계열 예측치 $\theta(t|t-1)$에 대한 작은 보정만이 일어난다. 역으로 만약 표본 분산이 작다면 $\theta(t|t)$는 최근의 표본 추정치 $Y(t)$와는 아주 다른 값이 될 것이다.

KF는 반복적인 방법으로 상태벡터 $S(t)$의 최소평균제곱오차를 제공하며, 또한 KF는 새로운 자료가 각 주기에서 이용될 수 있는 실시간 상황에 적합하다. 그러나 자료가 시간 t 이후에 이용될 수 있을 경우에는 추정치 $S(t|t)$는 이러한 새로운 정보를 반영시키지는 못할 것이다. 왜냐하면 KF는 시간 t의 앞쪽으로만 이동

하기 때문이다. 이 전 주기의 추정치들에 대한 부분적인 최적화는 평활을 통해 쉽게 조정될 수 있다.

평활의 방법은 두 가지 형태의 KF 추정량을 결합하는 방법이다. 첫 번째 형태의 KF 추정량은 이 전에 설명했던 것과 같이 전방 Filter 추정치로써 시간 t에서 모든 과거 표본자료와 현재 표본자료에 의해 추정되는 $S(t|t)$이다. 두 번째 형태의 KF 추정량은 후방 Filter 추정치로써, 표본 주기의 끝에서 출발하여(가령 $t=n$에서 출발) 처음까지 진행하며, 미래의 자료에만 근거하여, 각 시간 t에서의 예측치들을 이끌어 내는데 이러한 추정량을 $S(t|t+1)$로 나타낸다. 이 때 최적인 평활된 추정량은 두 추정량의 평균제곱오차 값들의 비율로 결합되어 만들어지는데 다음과 같다.

$$S(t|n) = P(t|n)\left\{ \frac{S(t|t)}{P(t|t)} + \frac{S(t|t+1)}{P(t|t+1)} \right\},$$

$$\text{단, } P(t|n) = 1/\left\{ \frac{1}{P(t|t)} + \frac{1}{P(t|t+1)} \right\}$$

한편, $S(t|n)$의 공분산 표현식으로부터 $\frac{1}{P(t|n)} = \frac{1}{P(t|t)} + \frac{1}{P(t|t+1)}$을 얻을 수 있고, $P(t|n) - P(t|t)$는 음 반정치(negative semidefinite)임을 알 수 있다. 따라서 $S(t)$의 평활된 추정량은 전방 Filter추정량보다는 훨씬 좋은 추정량이 되며, 이러한 이유 때문에 노동력 추정치들은 평활 알고리즘을 이용하여 만들어진다.

7.4 보충 설명

상태공간 형식은 signal성분을 세분화하는데 상당한 유연성을 허용하며, 특별한 경우들로써 표본조사 방법에 근거한 다음의 2가지 형태의 모형을 포함한다.

만약 $Q = Cov(v_\beta(t)) = 0$ 이고 $e(t)$와 $u(t)$가 백색잡음오차 변수이면, 시스템은 Ericksen(1974)의 표본회귀모형이 된다. 이 경우 signal 적출 문제는 관측된 표본자료에 대해 가중 최소제곱방정식을 적합시켜 해결된다.

$\beta(t) = 0$, $Q = 0$일 경우 Wiener-Kolmogorov의 signal 적출 이론에 근거한 모형을 얻는다. 회귀식의 평균은 없어지고 signal은 공분산 정상과정이 된다. 추가적으로 $e(t)$과정의 분산과 $ARMA$ 모수들이 상수인 경우에는 $e(t)$가 공분산 정상과정이 된다.

Scott와 Smith(1974)는 공분산이 추정되어야 하는 조사자료에 대해 전통적인 signal 적출 방법을 채택하였다. Bell과 Hillmer(1987a)는 signal 과정에서 비 정상성을 다루는 방법들을 토의하였다.

앞서 소개되었던 내용에서는 signal의 비 정상성 문제를 회귀변수들과 회귀계수들의 확률적인 변화로 다루었다. 회귀계수들의 움직임을 통제하는 (7.10)식의 추이방정식은 다양한 형태를 수용할 수 있다(Los(1985)). 이러한 회귀계수들은 독립인 랜덤워크를 따르는 것으로 명시한다(즉, $T_\beta = I$ 이고 Q가 대각행렬임을 명시).

CPS 자료를 이용한 연구는 많은 연구가 진행되고 있지는 않다. Hausman과 Watson(1985)은 CPS 4-8-4 연동표본교체와 합성절차를 통합시킨 전 지역의 십대의 실업률 시계열에 대한 오차과정의 $ARMA(1, 15)$모형을 개발했다. Bell과 Hillmer(1987b)는 Train(1978)등에 의해 추정된 design에 근거한 자기공분산의 근사로써 $ARMA(1, 1)$모형을 개발했다.

CPS와 같은 복잡한 조사에 대해서 공분산 추정치를 산출하는 것은 대규모의 자료가 사용되어 비용이 많이 든다. 최근에는 필요로 하는 모든 자료를 이용하지 않고 이러한 추정치를 산출하는 방법이 모색되고 있다.

설계에 근거한 표본오차 공분산 계산의 어려움 때문에 방정식에서의 오차들의 효과와 표본오차의 효과를 추정하지 않고 회귀방정식을 적합시키기도 한다. 만약

두 성분 오차들이 $ARMA$과정이면, 이 때 합도 $ARMA$과정이 된다(Granger and Morris(1975)). 즉,

$$u(t) \sim ARMA(p_u, q_u), \quad e(t) \sim ARMA(p_e, q_e) \text{ 이면}$$

$$w(t) = u(t) + e(t) \sim ARMA(p, q)$$

여기에서 $p \leq p_u + p_e$, $q \leq \max(p_u + q_e, p_u + q_e)$. KF는 회귀성분과 총 오차를 적출하기 위해 사용될 수 있다.

7.5 실업률 추정에 적용

CPS외의 자료로써 미 연방-주의 UI 체계로부터 생산된 실직보험가입자료(UID)와 비농인구에 대한 경상고용통계조사(CES)자료가 있다. 이러한 자료들은 1960년대 초 이래로 주의 특정 추정치를 생산하기 위하여 핸드북 방법에서 이용되었던 자료들이다.

UI자료와 CES자료로 설명되지 않는 순환적이며 계절적인 노동력의 움직임을 통제하기 위하여 표본오차의 영향을 줄이는 방법으로 선택된 CPS자료로부터 변수들이 구성된다. 이러한 방법은 구성된 변수들의 오차 문제를 다루기 때문에, 즉 계수들보다는 오히려 종속변수의 참값을 추정하는 데에 초점이 맞추어져 있기 때문에 기존의 방법과는 다르다.

설명변수들로 주의 특유한 CPS자료를 이용하는 것이 바람직하나 그렇게 하기 위해서는 변수들이 갖고 있는 오차를 명확하게 설명할 수 있는 모형이 필요하게 된다. 보험가입 실직자 수에 대한 월별 주단위 자료가 CPS와는 독립적으로 얻을 수 있는 실직자에 대한 최신정보이다. 이러한 자료가 실업률 모형 개발을 위한 출발점이 된다. 실직보험자료는 UI혜택을 신청하고 있는 노동자들의 수를 완전히 집계한다. 일반적으로, 해고되어 주의 특정한 재정상의 적격 기준을 충족시킨 노동자들

에게만 혜택이 돌아간다. 대조적으로 CPS에서 이용되는 실직의 개념은 조사기간 동안 직업을 갖고 있지 않은 모든 사람들, 해고되어 직업을 찾고 있는 사람 또는 일시 해고되어 대기발령 중인 사람들을 모두 포함한다. 미국의 경우 실직보험자료에서 실직자에 포함되지 않은 실직그룹들을 살펴보면 <표11>과 같다.

<표11> 실직보험자료(UID)에서 집계 안된 비고용 그룹

1. 다음의 범주에 해당되는 실직자
   a. Exhaustees : 수혜 권리를 다 써버린 노동자들
   b. 재정적 부적격자 : 주의 적격 요구기준을 충족시키지 못하는 우선 고용자 또는 소득을 갖고 있는 노동자들
   c. 유예된 비 신청자 : 실직시기의 처음에 혜택을 신청하지 않은 적격한 노동자들
2. 노동력 신규 유입자 : 최근의 실직기간 전에 노동인구에 편입되지 않았던 노동자들
3. 이직자 : 이 전의 직업을 그만두고 다른 직업을 찾고 있는 노동자

UI가 집계 않는 부분들 중에서 신규 유입자가 차지하는 비율이 가장 크며, 다니는 직장을 그만두고 다른 직업을 찾는 이직 희망자들은 적어도 그 기간만큼은 보험상의 혜택을 받지는 못한다. 만약 UI혜택을 받지 못하는 실직자의 상대적인 규모가 시간에 따라 안정적이라면, 보험 지불요구율은 전체 실업률 추정에 대용이 될 수 있다. 실제로 노동시장에서 특히 실직자와 신규 유입자 간의 실업률 분포는 순환적이며 계절적인 특성변화를 나타낸다.

먼저 실업률 분포에 있어서 계절적 변화를 고려해 보자. 가장 중요한 현상은 실직자와 유입자는 매우 다른 계절적 형태를 띤다는 점이다. 청년과 여성이 유입자의 대부분을 차지하고, 청년층의 실업은 휘발성의 계절적 형태를 띠는데, 이는 학기가 끝나면서 신규 노동력이 유입되고 학기 시작과 함께 빠져나가는 청년층의 노동력의 변화가 반영되었기 때문이다. 이와는 대조적으로 성인 남성의 실직에 있어서 가장

일반적인 원인은 자동차 산업과 건설 산업과 같이 산업의 연간 생산 순환주기에 영향을 받아 계절적인 해고와 재 고용이 발생한다. <표12>는 CPS 신규유입 및 실직률의 계절적 형태(40개 주에 대한 평균값)와 전체 비율에 대한 이러한 항들의 순수 효과 간의 차이를 나타내며, 또한 실직보험 가입자에 대한 계절적인 형태를 보여준다.

여기에서 100보다 큰 값은 평균실업률보다 큰 달을 말한다. 신규 유입률은 겨울에는 평균실업률보다 낮고 여름에는 평균실업률보다 높다. 반면, 실직률은 반대의 형태를 보인다. 각 그룹에서의 큰 계절 실업률은 전체 실업률에 강한 영향을 미친다. 실직자와 신규 유입자는 경기순환과 다른 형태를 띤다는 점에서 수치적으로 중요하며, 전체 실업률의 약 반 정도를 설명한다. 이직자는 전체의 약 15%를 설명하므로 수치적으로 덜 중요하다. 늦여름과 가을에 구별되는 계절적 형태를 나타낸다. 실직보험 지불요구율은 실직자의 계절적 형태를 따르나 신규 유입 또는 이직자의 계절적 형태를 따르지는 않는다.

<표12> 40개 주에 대한 실업 계절 요인들에 대한 평균(1979-85)

월	CPS 실업률				UI 지불요구율
	전체	신규유입자	실직자	이직자	
1월	110.6	99.1	120.9	104.3	131.2
2월	110.9	98.1	126.1	100.7	133.8
3월	105.3	96.2	115.7	95.5	120.8
4월	97.0	90.7	103.5	91.3	104.0
5월	93.2	96.4	91.0	92.8	91.0
6월	106.0	127.2	89.6	93.6	85.8
7월	98.7	106.0	90.8	101.0	95.9
8월	98.5	102.3	92.1	110.7	90.2
9월	95.5	102.7	84.7	112.8	77.6
10월	93.6	98.3	88.0	107.0	79.3
11월	94.9	94.5	93.5	101.3	87.7
12월	95.9	88.4	104.1	89.5	102.4

* <표12>에서 계절 요인들은 X-11로 계산됨. CPS 실업률의 분모는 각각의 범주에 대해서 CPS 고용과 실직인원의 합이고, UI 지불요구율에서 분모는 CES고용의 총계임.

다음의 <표13>은 경기순환과 관련한 실직률의 분포의 변화를 나타낸다. 경기후퇴 기간 중에는 노동력 요구는 떨어지고 실직자는 증가한다. 따라서 실직보험 지불요구율은 충분히 순환적인 지표가 된다. 그러나 실직보험 지불요구는 실직자의 순환적인 변화를 전적으로 반영하지는 못한다. 경기후퇴의 후반부 쪽에서는 실직기간이 길어져서 UI 혜택을 모두 써버린 노동자들이 증가한다. 또한 일단 재 고용되면 이러한 노동자들이 차후의 실직기간 동안 수혜에 대한 자격을 얻기 위한 고용신뢰도와 충분한 임금을 얻기까지는 얼마간의 시간이 걸린다. 중요한 사실은 실직자들에 대한 UI 범위를 살펴보면 오랜 기간에 걸쳐 계속해서 줄어든 적이 있었다는 점이다. <표13>에서 이러한 사실을 확인할 수 있다. 경기후퇴 기간 동안 증가하는 대신 UI 값이 줄어들었다.

Burtless et al.(1984)와 Corson et al.(1988)의 연구에 의하면 이러한 현상은 경제적인 변화나 인구 통계적인 변화와는 무관하며 UI 대상자들이 혜택을 적용 받지 못한 이유 때문으로 실명한다. 정책의 변화가 주요한 원인으로 주목된다.

이상의 토의는 대표적인 주의 움직임을 설명한 반면, 모형화 과정에서 실명되어야만 하는 중요한 주 간의 차이점들이 존재한다. UI 자료에서 보면 주의 적격요구기준, 수혜기간, UI 적용범위에 대한 행정관례에서의 변동사항들이다. 순환적이며 계절적인 움직임에 있어서 실제적인 차이들은 회귀계수와 모형의 모수에 영향을 끼친다.

<표13> 실직범주들의 상대적인 크기(40개 주 평균)

Year	U.S 실업률	Percent of Total CPS Unempoyment				
		CPS실직자 UI지불요구 (%)	지불요구	실직자	신규유입자	이직자
76	7.7	93.1	36.9	43.1	41.9	14.8
77	7.1	86.5	34.5	41.3	43.6	14.9
78	6.1	87.6	31.4	37.7	46.1	16.2
79	5.8	87.5	32.8	39.4	44.4	16.1
80	7.1	76.7	35.8	47.6	38.3	14.0
81	7.6	68.1	31.9	48.8	38.4	12.7
82	9.7	61.8	34.3	56.3	34.5	9.1
83	9.6	50.7	27.7	55.1	35.5	9.2
84	7.5	52.2	25.7	50.1	38.9	10.9
85	7.2	56.6	26.9	48.3	40.1	11.5
86	7.0	60.3	27.9	47.7	38.9	13.3
87	6.2	57.0	25.3	45.6	40.2	14.0
88	5.5	62.3	26.4	43.9	40.4	15.5

UI 적용에 있어서 변화는 <표14>에 주어졌다. <표14>는 주별 전체 CPS실업자에 대한 UI 지불요구율의 연간 평균값들의 전체평균(%), 주 내에서의 CV값, 연간 평균값들 중 최소값과 최대값을 나타냈다.

<표14> 주별 전체 CPS 실업자에 대한 UI 지불요구자(1976-87)

주	UI지불 요구(%)	CV	최소값	최대값	주	UI지불 요구(%)	CV	최소값	최대값
AL	25.5	23.7	17.5	37.5	MT	32.2	21.6	23.1	44.2
AK	57.1	21.2	43.1	83.0	NE	30.0	18.7	20.7	43.9
AZ	24.3	13.8	19.8	29.5	NV	34.7	19.7	25.3	48.2
AR	28.1	21.6	20.2	38.5	NH	26.4	17.4	19.2	34.7
CO	23.6	10.9	20.2	28.1	NM	24.2	10.5	20.4	28.1
CT	37.5	17.0	29.3	48.9	ND	31.1	14.7	24.3	36.7
DE	29.9	17.1	22.8	38.0	OK	27.5	21.7	19.4	39.0
DC	26.6	10.7	21.9	30.3	OR	37.6	16.7	28.8	48.4
GA	25.8	15.0	21.2	33.4	RI	51.2	12.7	39.5	61.4
HI	32.9	9.5	29.0	39.8	SC	26.3	18.1	19.8	33.9
ID	31.6	15.4	24.7	39.5	SD	22.5	30.6	14.7	35.0
IN	24.4	17.5	19.4	33.5	TN	27.2	26.2	18.1	41.0
IA	30.3	20.9	22.6	42.1	UT	31.4	25.7	19.9	41.1
KS	35.5	13.6	29.5	45.9	VT	40.0	9.7	33.9	46.3
KY	29.2	29.7	17.2	40.0	VA	17.6	20.1	13.3	24.7
LA	27.9	20.6	20.0	37.6	WA	36.1	15.7	30.1	49.1
ME	35.8	11.2	28.7	41.1	WV	32.8	24.7	21.1	42.7
MD	28.5	12.2	23.6	34.0	WI	34.2	24.1	24.8	47.1
MN	33.7	18.7	24.0	44.4	WY	28.7	28.7	20.7	49.7
MS	26.1	19.7	20.1	35.3	ALL	31.0	23.2	17.6	57.1
MO	32.8	19.5	24.3	44.2					

비율모형의 회귀성분의 일반적인 형태는 다음과 같이 주어진다.

$$\text{실업률} = \text{Intercept (t)} + \beta_1(t)\ (\text{지불요구율}) + \beta_2(t)\ (\text{인구 대 고용비(EPR)}) + \beta_3(t)\ (\text{신규유입률}) \tag{7.16}$$

여기에서 $\text{지불요구율} = \frac{contiued\ claims\ w/o\ earnings}{CES\ employment} \times 100$,

$$\text{인구 대 고용비} = \frac{CPS\ employment}{CPS\ 16+\ population} \times 100\ ,$$

$$\text{신규유입률} = \frac{CPS\ \text{고용 신규유입자}}{CPS\ \text{신규유입자} + CPS\ employment} \times 100\ .$$

지불요구율은 UI 혜택을 받고 있는 실직자들의 상대적인 규모를 나타내는 측도이다. 인구 대 고용비는 보험 지불요구 인원에 포함되지 않은 실직자들을 나타내는 측도이다. 숙련된 노동자들에 대한 상대적으로 고정된 노동력 관계비가 주어진다면, 경기순환 동안의 그들의 실직은 역으로 인구 대 고용비와 관계가 될 것이다. 이직자의 수에 영향을 미치는 노동요구에 있어서의 계절적 변동을 인구 대 고용비에서 찾아낼 수 있다. 즉 여름에 정점을 취하고 가을에 떨어지는 계절적 변동을 보인다. 이러한 주기 동안 수많은 계절적 직업들이 생기며, 이직에 기인한 실직이 증가한다.

대부분의 주에서는 고용 대 인구비의 분모의 값을 CPS 고용자료를 이용하여 계산한다. 주의 CPS 고용자료는 표본추출오차의 영향을 받을 가능성이 있지만, CV의 값은 주의 CPS 추정치 보다는 5~6배정도 작다. 소수의 주에서 인구 대 고용비의 계산에서 CES자료를 이용한다. 대부분의 주들에 대해서 CES는 CPS 고용측도와는 어느 정도 다른 계절적 형태를 갖고 있으며 실직률의 참값과는 높은 상관성을 갖고 있지는 않은 것 같다. 표본추출오차의 효과를 줄이기 위하여 CPS 신규 유입률 변수는 주보다는 훨씬 큰 지리학적 지역(전국 또는 센서스 지역)에 대해서 계산된다. 몇몇 경우에 있어서는 주의 CPS 신규 유입률의 3-month 이동평균이 사용된다.

40개 주의 각각에 대한 모형들이 1976-87년 동안의 월별 CPS 실직률의 시계열 자료에 적합되었다. 전에 설명한 것과 같이 CPS 표본자료는 다음과 같이 signal성분과 noise성분으로 표현된다.

$$Y(t) = \theta(t) + e(t) \ ,$$

$$\text{단,} \quad \theta(t) = X(t)\,\beta(t) + u(t).$$

여기에서 계수들은 다음과 같은 랜덤워크를 따른다.

$$\beta(t) = \beta(t-1) + v_\beta(t)$$

$u(t)$와 $e(t)$의 효과는 분리하여 추정하지 않기 때문에 관측된 값들은 다음과 같이 표현할 수 있다.

$$Y(t) = X(t)\,\beta(t) + w(t) \quad ,$$

$$\text{단,} \quad w(t) = u(t) + e(t) \sim ARMA(p\,,\ q)$$

$S_w(t)$를 $\max(p\,,\ q+1)$과 같은 계(order)를 갖는 $w(t)$의 상태벡터라 하면, 상태공간 모형은 다음과 같은 전이방정식(transition equation)을 갖는다.

$$S(t) = \begin{bmatrix} \beta(t) \\ S_w(t) \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & T_w \end{bmatrix} \begin{bmatrix} \beta(t-1) \\ S_w(t-1) \end{bmatrix} + \begin{bmatrix} I & 0 \\ 0 & \Gamma_w \end{bmatrix} \begin{bmatrix} v_\beta(t) \\ v_w(t) \end{bmatrix} \ ,$$

$$Y(t) = [\,X(t)\ 1\ 0\ \cdots\ 0\,]\,S(t) \ .$$

이 시스템의 모수들은 다음과 같다.

$$Cov(v_\beta(t)) = Diag(\sigma_{\beta_1\beta_1}\ ,\ \cdots\ ,\ \sigma_{\beta_k\beta_k}) \ ,$$

$$Var(v_w(t)) \ ,$$

T_w는 p AR 모수들을 갖고, Γ_w는 q MA 모수들을 갖는다.

이러한 모수들은 우도함수의 새로운 형태를 이용하여 추정된다 (Schweppe(1965)). 만약 백색잡음오차 $v_\beta(t)$와 $v_w(t)$가 정규분포를 따른다면, 1-step 앞의 예측오차들인 $\hat{Y}(t)$는 독립인 $N(0, f(t|t-1))$을 따르는 확률변수가 된다.

관측된 표본오차의 결합확률은 각각의 밀도함수의 곱으로 표현된다. 만약 상태벡터가 l개의 비 정상원소를 갖고 있다면, 결합밀도함수는 처음 l개의 관측점에 대한 조건부 함수로 표현되며, 미지인 모수들인 Ω의 함수로써 로그우도함수

$$L(\Omega) = -\frac{1}{2}\left\{\sum_{t=l}^{n}\log f(t|t-1) + \frac{Y(t)^2}{f(t)}\right\}$$

는 상수의 범위내에 있게된다. 만약 Ω가 주어진다면, 초기값 $S(l+1|l)$, $P(l+1|l)$을 계산하기 위한 처음의 l개의 관측값을 이용하여 $L(\Omega)$를 결정하는데, 이때 KF반복법이 사용된다. 일반적으로 이러한 방법은 어려운 비선형 최적화 문제에 해당된다. 초기에 우리는 $Cov(v_\beta(t)) = qD$ 로 나타냈다. 여기에서 q는 상수이고, D는 사전에 명시된 상수들의 대각행렬이다. 또한 $w(t)$에 대한 1계인 AR모형으로 출발하여, 두 모수들의 추정문제를 다루었다. 계수들과 AR 모수 값의 변동의 정도에 대한 개략적인 추정치들을 계산하였고, 몇몇 경우에서 이러한 추정치들은 Watson과 Engle(1983)에 의해 개발된 EM-scoring 알고리즘의 초기값들로 이용되어 좀 더 다듬어지게 되었다. 이러한 알고리즘은 계수의 변화와 관측오차를 동시에 고려하는 일반적인 자기회귀 구조에도 이용된다.

모수들에 대한 초기값들은 합리적인 반복횟수 내에서 수렴값을 얻기 위해 매우 중요하다. 계수들의 표준편차($\sqrt{\sigma_{\beta_i\beta_i}}$)는 관측오차의 표준편차의 약 0.6%정도였다. 초기값들로 이러한 사실들을 이용한다면 EM 알고리즘은 일반적으로 3~6회의 반복 후 수렴하게 된다. Harvey와 Phillips(1979), Ansley와 Kohn(1985), Bell과 Hillmer(1987a)에 의하면 KF를 초기화하는 많은 방법들이 있다.

8. 기타 외국의 소지역 추정법 활용 현황

8.1 영 국

영국은 분기별로 노동력을 조사하여 실업 통계를 국가 수준에서 작성 발표하였으나 유럽 연맹에서 실업 통계를 표준화할 뿐 만 아니라, 지방자치정부 등에서 실업 통계의 소요를 제기함에 따라 ONS(Office for National Statistics)와 Southampton 대학이 협동 연구로 소지역 추정법을 적용한 실업 통계 작성에 관한 프로젝트가 추진중이다.

영국의 통계 작성 단위는 LAD(Local Authority District)가 우리 나라 시군구 정도 단위이고, 그 하부 단위는 ward가 있으며 ward단위까지 실업 통계 작성이 요구되고 있다. 영국에는 408개의 LAD가 있으며 규모도 16세 이상 인구가 7천에서 76만 7천 명까지 범위가 크다.

영국에서 연구되는 소지역 추정 모형은 two-level 변환 모형으로 기본적인 형태는 다음과 같다.

$$Y_{ij} = \beta_{0j} + X_{ij}^{T}\beta_{ij} + \varepsilon_{ij} \quad ; \; i = 1, \cdots, m_j \; , \; j = 1, \cdots, n \qquad (8.1)$$

여기에서 $\varepsilon_{ij} \sim (0, \sigma_{\varepsilon}^2)$ 으로 오차항이고, $\beta_j = (\beta_{0j}, \beta_{ij}^t)^t \sim N(\beta, \sum\beta)$는 소지역 effect를 고려한 random 성분, $Y_{ij} = g(z_{ij})$는 원천적 반응 변수 Z_{ij}의 변환된 변수이다.

이 문제에서 핵심적인 것은 변환 $g(\cdot)$이며, 회귀모형분석 등에서 이용되는 package등을 통해서 구해질 수 있고, 이 함수가 단조이면 역함수 $h(\cdot) = g^{-1}(\cdot)$를 찾을 수 있다.

간단한 예로써 $g(Z) = \log(Z)$인 경우를 가정하면 $E(Z_{ij})$의 추정량은 $\hat{Z}_{ij} = \exp(\hat{Y}_{ij})$로 주어지고 $\hat{Z}_{ij}$의 표준 오차의 추정값은 다음과 같다.

$$\hat{d}_{ij} = \{\widehat{Var}(\hat{Z}_{ij})\}^{1/2} \approx \exp\{\hat{E}(y_{ij})/2\}\{\widehat{Var}(\hat{Y}_{ij})\}^{1/2} \qquad (8.2)$$

$$= \exp\{\hat{Y}_{ij}/2\}\ \hat{\sigma}_{\hat{Y}_{ij}}$$

실제로 경활조사를 이용한 소지역 실업자 추정을 소개하기로 한다. 경활 조사는 분기별로 6만 가구를 표본으로 선정하여 조사하는데 표본 조사구는 5개 분기를 연속 조사 후 연동 교체 표본으로 대체된다. 실업자의 특성 파악을 위해서 LAD를 6개 범주(성별-연령대별)로 분류하고 실업 보험 신청 자료를 보조 정보로 활용한다.

Multi-level 변환 모형을 적용하기 위한 변수들은 다음과 같이 정의한다.

$Z_{ij} = \dfrac{ILO_{ij}}{POP_j} = V_{ij}P_{ij}$,

$V_{ij} = ILO_{ij}/POP_{ij}$; j 소지역 i 범주에서 실업률 ,

$P_{ij} = POP_{ij}/POP_j$; j 소지역에서 (i,j) cell의 구성비 ,

$Y_{ij} = \log(Z_{ij})$; $Z_{ij} > 0$,

$X_{1ij} = \log$(실업 보험 신청자 구성비) ; 성별-연령대 구분 ,

$X_{2ij} =$ 기타 Covariates .

실제 자료를 대입하여 적합한 multi-level 변환 모형은 아래와 같이 주어졌다.

$$\hat{Y}_{ij} = -2.117 + (0.582 + \hat{V}_j)X_{1ij} - 1.561S_{ij} + 0.271a1_{ij} \qquad (8.3)$$

$$+ 1.516a2_{ij} + 0.896a2_{ij}*S_{ij} - 0.270S_{ij}*X1_{ij} + 0.366a2_{ij}*X1_{ij}$$

여기에서 $a1_{ij} = \begin{cases} 1 & i\text{범주가 첫째 연령 범주일 때}(16-25) \\ 0 & \text{그 외} \end{cases}$,

$a2_{ij}=\begin{cases} 1 & i\text{범주가 두번째 연령 범주일 때}(26-49) \\ 0 & \text{그 외} \end{cases}$,

S_{ij} : 성별 표시 ,

$\hat{V}_j$: j소지역의 *random*성분 ($\sigma_v^2=0.00071$).

위의 적합 모형은 모집단 자료를 이용해서 분석한 결과로서 변동의 72%까지 설명함을 보였다. ($R^2=72\%$)

8.2 캐나다

캐나다는 노동력 조사를 매월 53,000 가구를 추출하기 위해서 SRU (Self-Representing Units)와 NSRU (Non Self-Representing Units)로 구분하여 2단 층화 추출법을 사용하고 있다. 1차 단계에는 2~300 가구를 하나의 집락 단위로 하여 PPS 추출법으로 집락을 뽑고, 2 단계 추출에서는 계통 추출법으로 6~10개 가구를 선정하여 표본 가구는 6개월간 조사된 후에 교체하는 연동 교체 표본 관리를 적용하고 있다.

캐나다는 53개 취업보험지역 (EIR : Employment Insurance Regions)으로 구분하여 각 EIR별로 실업 보험 혜택과 기간을 결정하는데 실업률을 기준으로 하므로 EIR별로 신뢰성있는 실업률의 추정이 요구되어 소지역 추정법에 대한 연구가 오랫동안 진행되었다.

초기에는 실업자와 취업자 추정을 위한 복합 추정량과 표본 크기 의존 추정량 (SSD : Sample-size Dependent)을 고려하였다.

복합 추정량은 사후 층화 추정량과 합성 추정량의 일차 결합으로 나타냈으며 추정량과 적정 가중값은 아래와 같다.

$$\hat{Y}_a=\alpha\ \hat{Y}_{a\cdot p}+(1-\alpha)\ \hat{Y}_{a\cdot s} \qquad (8.4)$$

$\hat{Y}_{a\cdot p}$는 a 소지역의 사후 층화 추정량이고, $\hat{Y}_{a\cdot s}$는 a 소지역의 합성 추정량이며 α 의 최적값은 다음과 같다.

$$\alpha^* = \frac{mse(\hat{Y}_{a\cdot s}) - E(\hat{Y}_{a\cdot s} - Y_a)(\hat{Y}_{a\cdot P} - Y_a)}{mse(\hat{Y}_{a\cdot s}) + mse(\hat{Y}_{a\cdot p}) - 2E(\hat{Y}_{a\cdot s} - Y_a)(\hat{Y}_{a\cdot p} - Y_a)}$$
$$= \frac{mse(\hat{Y}_{a\cdot s})}{mse(\hat{Y}_{a\cdot s}) + mse(\hat{Y}_{a\cdot p})}$$

혼합 추정량을 이용시에는 mse의 참값을 구할 수 없으므로, 조사된 표본 자료 또는 기타 행정보고 자료를 통해서 $mse(\hat{Y}_{a\cdot p})$와 $mse(\hat{Y}_{a\cdot s})$를 추정하지만 계산 절차가 복잡하고 가중값이 표본 자료에 따라 불안정한 경우에는 복합 추정량의 변동폭이 커지는 문제점이 있다.

소지역의 표본의 크기에 따라 사후 층화 추정량의 안정성이 다르기 때문에 합성 추정량의 반영도를 표본 크기에 의존하도록 복합 추정량 형식을 변화하였다. 표본의 크기는 쉽게 알 수 있으므로 계산이 간편하고 조사된 표본의 결과를 잘 반영한 추정량으로 인식되어 현재 대부분의 소지역의 노동력 통계 생산에 적용되고 있다. a 소지역의 실업 통계의 추정량은 아래와 같다.

$$\hat{Y}_a = \sum_g \left[\lambda_{ag} \frac{N_{ag}}{\hat{N}_{ag}} \hat{Y}_{ag\cdot p} + (1-\lambda_{ag}) \frac{N_{ag}}{N_g} \hat{Y}_{g\cdot s} \right] \qquad (8.5)$$

여기에서 a는 소지역을 뜻하고 h는 층을 나타내며, g는 지역(시지역 또는 농촌지역)을 나타낸다. $\hat{Y}_{ag\cdot p} = \sum_h \hat{Y}_{ag\cdot p}$ 이고, $\hat{Y}_{g\cdot s} = \sum_h \hat{Y}_{gh\cdot s}$이며 N은 모집단의 크기를 의미한다. a 소지역의 표본 크기가 충분히 크면 합성 추정량 ($\hat{Y}_{g\cdot s}$)는 무시된다. 즉 $\lambda_{ag} = 1$이 된다. λ_{ag}는 아래와 같이 정의되며,

$$\lambda_{ag} = \begin{cases} 1 & \text{만일 } \hat{N}_{ag} \geq N_{ag}/k \\ k\,\hat{N}_{ag}/N_g & \text{(그 외 경우)} \end{cases}$$

k는 소지역에서 표본의 크기가 적음에 따라 합성 추정량의 성분을 반영할 것인가

를 판단하는 숫자이다. 만일에 단순임의 추출에서 $k=2$ 로 놓으면

$$\lambda_{ag}=\begin{cases} 1 & \text{만일 } n_{ag}\geq E(n_{ag})/2 \\ 2n_a/E(n_{ag}) & \text{(그 외 경우)} \end{cases}$$

와 같으므로 실제 표본 크기가 기대 표본 크기의 절반 이하이면 합성 추정량을 반영하게 된다는 의미이다. 표본 크기 의존 추정량은 각 소지역 별로 적정 k 값을 정해야 하므로 복잡하기 때문에 고정된 k 값을 사용하고 있으나 소지역의 특성을 제대로 반영하지 못하는 문제점이 있다.

8.3 프랑스

노동력 조사로부터 국가 수준의 노동 통계는 일년에 한 번씩 발표하여 이용자들에게 불편을 주었으나, 지역에서 실업 통계를 생산할 때 일정한 오차 범위 내로 기준을 충족하도록 하는 소지역 추정법에 관한 연구가 진행되고 있다.

표본 설계는 지역에 의해서 층화하였으나 실업자 수 또는 실업률 등의 집적된 자료의 이용시에는 무응답을 보정하여 국가 수준으로 통계를 작성한 후에 성별-연령으로 분할했으므로 지역 통계의 신뢰성에 대해서 유의해야 할 것이다.

프랑스 통계청에서 취업과 실업에 관한 지역 통계의 생산을 위해서 노동력 조사와 센서스 뿐 만 아니라 실업 보험 신청자료 등 행정 업무 자료를 이용하는 체계를 발전시켜왔으나 소지역이나 세분화된 범주의 통계 작성은 미흡하다. 특히 취업 통계는 연말에 각 회사로부터 취업자 수에 대한 자료를 수집하고 있으나 작년말 기준으로 변화율에 대한 자료로 활용하여 국가 수준에서 비율에 대한 통계를 작성하고 있으며 실업자의 수에 대한 직접 통계를 작성하지는 않는다. 그래서 취업자 통계는 실업 보험과 센서스 자료 등이 핵심적인 역할을 하고 있다. 취업 통계는 해당 익년에 이용가능하다.

실업 통계는 분기별로 지역 통계와 국가 통계를 동시에 발표하며, 주로 이용되는

자료는 노동력 조사와 구직 등록자의 수이다. ILO 기준의 실업자와 실제 구직 등록자 및 노동력 조사에서 실업자와 차이를 반영하기 위해서 노동력 조사에서 추정된 ILO 실업자와 구직 등록자 수의 국가적 비율을 추정하여 활용하고 있다.

경제활동인구는 실업자와 취업자를 합해서 추정하지만 취업자는 직장을 갖고 있는 사람과 가사를 돌보는 사람을 합산해야 하므로 매 해마다 연말에 외삽법으로 조정하여 분기별 통계를 수정하여 시계열 통계로 관리한다.

취업자 통계에서 문제점은 연말에 센서스 자료를 이용하고 있으나 센서스 간격이 너무 길어서 인구 변동과 상황 변화를 제대로 반영할 수 없다는 점이다 (센서스:1968, 1975, 1982, 1990, 1999). 특히, 지역 취업 통계 작성 시에는 더 심각해질 수 있다.

실업 통계 작성에서 문제점은 노동력 조사에서 국가 단위의 ILO 실업 통계는 작성할 수 있으나 지역 단위 또는 소지역 단위의 실업 통계는 생산할 수 없으며, 구조적인 특성(인구 사회학적 구성)이 각 지역이 국가와 동일하다는 전제에서 계산되며 표준 오차를 계산하지 못하고 있다.

취업 통계 작성의 취약점을 보완하기 위해서 "ESTEL"이라는 프로젝트가 수년간 추진 중에 있으나 아직 1996년 통계를 시험 중에 있다. 특징은 국가와 지역의 취업 통계의 질과 신뢰도를 제고하는데 있으며 매년 말을 기준으로 하여 조사 자료와 행정 자료를 반영하고 있다. 센서스 자료를 기준으로 반영하는 것은 센서스 해에만 적용하고 다른 해에는 연말의 행정 업무와 조사 자료를 기준으로 삼는 것이 큰 차이점이다.

핵심적인 과제는 경제 활동 관련 통계의 기준은 센서스 정보이므로 센서스 실시를 일정한 간격으로 실시하는 계획일 것이다. 매 5년마다 센서스에 준하는 대규모 통계 조사 계획을 추진하고 있다.

9. 시 · 군 · 구의 실업통계 개발 예

9.1 개 요

충북의 행정구역은 2구 2시 8군으로 편성되어 있다. 다음의 <표15>는 1999년 4월의 경제활동인구 조사 결과를 요약한 것이다.

<표15> 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

<표15>에 따르면 충북의 경제활동 참가율이 남자의 경우, 여자에 비해 비교적 높은 것으로 나타나고 있다. 다음의 <표16>은 충북지역의 1999년 4월 시와 군 지역을 구분하여 성별에 따른 경제활동 인구와 조사구 수의 현황을 나타낸 것이다.

9.2 시군구 실업자 추정

9.2.1 직접 추정량

경활조사 자료를 이용하여 시군구 실업자를 추정할 수 있는 방법으로써 앞에서 3가지를 언급하였으며 이 중에서 직접추정량은 사전에 계산한 동부와 읍면부에 대한 승수(통계청(1996), pp12)와 경활조사 자료만을 이용하여 식(2.1)을 통해서 계산

하였다. 또한 분산은 현재 통계청에서 사용하고 있는 연속 차의 분산공식을 적용하여 계산하였으며, 시군구별 실업자 추정값과 분산은 <표17>에 주어졌다. 특기 사항은 진천군의 표본조사구는 1개이므로 분산을 계산할 수 없어서 일반화 분산함수(GVF:Generalized Variance Function)(Wolter(1985), pp210)를 통해서 계산하였다.

<표16> 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

9.2.2 합성 추정량

여기서는 "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 소지역별로 실업자를 다음과 같이 추정한다.

1. 성별-연령대별 구성비를 이용한 시군구별 실업자 추정

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

$$\hat{y}_{i.} = \sum_{j=1} x_{ij}\ r_{.j}^{a} \qquad (9.1)$$

단, i는 시군구를 나타내고, j는 성별-연령대별 범주를 나타낸다. x_{ij}는 i 시군구의 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 계산할 수 있다. 즉,

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

의 관계로부터 구할 수 있다. 이를 통해 계산된 1999년 4월의 충북지역의 특성별 실업률은 <표18>과 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.

$$\hat{y}_{i.} = \sum_{j=1} x_{ij}\, r_{.j}^{\,b} \tag{9.2}$$

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

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

의 관계로부터 계산할 수 있으며, 그 결과는 <표18>과 같다. 식(9.1)과 (9.2)로 주어진 추정량의 분산은 각 범주별 실업률을 상수로 가정하고 x_{ij}의 일차결합의 분산식을 통해서 계산하였다.

9.2.3 복합 추정량

직접 추정량은 표본 수가 적기 때문에 분산이 클 뿐 만 아니라 표본의 크기에 민감하게 변동하고 합성 추정량은 각 범주들이 모든 시군구에서 동일하다는 전제조건이 어긋날 경우에는 편향이 커지기 때문에 두 가지 추정량의 문제점을 완화하기 위해서 두 추정값의 가중 평균 형식인 복합 추정량을 생각하게 되었다.

$$\hat{Y}_{i.}^{\,c} = w_i \;\; \hat{Y}_{i.}^{\,d} + (1-w_i) \;\; \hat{Y}_{i.}^{\,s} \tag{9.3}$$

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 추

정한 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 $mse(\hat{Y}_{i.}^{s})$의 계산은 복잡하고 어려우므로 $Var(\hat{Y}_{i.}^{s})$로 대체하여 w_i의 근사값을 다음과 같이 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(8.3)에서 주어진 추정량의 분산은 근사적으로 아래 식으로 계산할 수 있다.

$$\widehat{Var}(\hat{Y}_{i.}^{c}) = \frac{\widehat{Var}(\hat{Y}_{i.}^{s}) \cdot \widehat{Var}(\hat{Y}_{i.}^{d})}{\widehat{Var}(\hat{Y}_{i.}^{s}) + \widehat{Var}(\hat{Y}_{i.}^{d})} \qquad (9.4)$$

9.2.4 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 실업자 추정값을 다음과 같이 조정한다. 즉, 군지역의 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.

9.3 추정결과

<표17>에서 분산의 특징을 살펴보면 직접추정량의 분산이 가장 크게 나타났다. 합성추정량의 분산은 편향을 생략하였기 때문에 직접추정량보다는 작을 것으로 예상 했지만, 13배에서 200배 정도까지 감소하였다. 표본의 크기가 작을 경우에는 합성추정법이 직접추정법에 비해서 월등하게 효과적이며, 복합추정량이 직접추정량에 비해 보다 안정적인 특성을 보이고 있다. 3가지 추정법 중에서는 복합추정법이 다른 추정법에 비해 보다 효과적임을 수치적인 사례로 보여주고 있다.

<표17> 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

<표18> 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

10. 결 언

지방 자치제의 정착으로 각 시군구의 경제 활동 관련 통계의 필요성이 제기되었을 뿐 만 아니라 실업 대책 및 공공근로사업의 예산 배분 등으로 적절한 예산 배정의 기준 설정을 위해서 시군구의 경제활동인구, 취업자 및 실업자 수에 대한 국회와 행자부의 요청이 있었으나 합당한 방법이 없었다. 그래서 몇몇 시도에서 자체적인 통계 조사 계획으로 시군구 실업자 수를 추진하고자 시도하였으나 신뢰성있는 결과를 얻을 수 없었다.

시군구의 실업자 및 경제활동관련 통계는 필수적으로 작성되어야 하므로 많은 예산과 인원을 투입하지 않고도 시군구의 통계 생산이 가능한 소지역 추정법에 대한 연구는 시급하게 추진되어야할 것이다.

소지역 추정법은 미국이나 캐나다에서는 오래 전부터 활용되어 왔으며 영국이나

프랑스 등 유럽에서는 국가적인 차원에서 연구가 진행중이다.

본 연구에서는 지금까지 연구되어 활용되고 있는 내용을 중심으로 소지역 추정법을 요약했고 충북 시군구의 실업자 추정에 합성과 복합 추정법을 적용시켜 본 결과에서 안정적이고 손쉽게 시군구의 실업통계를 생산할 수 있음을 보여 주고 있다.

본 연구는 이제 시작일 뿐이다. 좀더 신뢰성있고 광범위한 정보(실업 보험 신청 dB, 구직 등록 dB)를 이용할 수 있는 방안과 경활 조사가 매월 패널 조사로 이루어지고 있으므로 이들의 시계열 특성을 이용하여 각 지역별로 model을 추정하여 베이지안 추정법의 적용 방안에 대한 연구를 추진해야만 2000년 센서스 후 경활 조사 표본을 개편할 때 소지역 추정법을 적용할 수 있을 것이다.

앞으로 정보화 시대에서는 정부 부처간의 유관 정보는 공유할 수 있는 체제도 함께 추진해야만 신뢰성있는 정부 통계를 생산하는데 도움을 줄뿐 만 아니라 점점 어려워지는 조사 환경에서 신속 정확한 통계의 생산에도 도움이 될 것이다.

참 고 문 헌

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

[2] 통계청(1997) 「표본개편 결과 보고서」

[3] Battese,G.E., and Fuller,W.A.(1981) Prediction of county crop areas using survery and satellite data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 500-505

[4] Battese,G.E., Harter,R.M., and Fuller,W.A.(1988) An error components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36

[5] Bureau of the Census (1978) The current population survey: *Design and Methodology*, Technical paper 40, Washington, D.C.: Richard Tiller

[6] Burtless, G., and Vroman, W.(1984) The performance of unemployment insurance since 1979, *paper presented at the December 1979 meeting of the Industrial Relations Research Association*, Dallas, Texas

[7] Corson, W., and Nicholson, W.(1988) An examination of declining UI Claims during the 1980s, *Mathematica Policy Research Draft Report*

[8] 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

[9] 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

[10] Datta, G. S., and Ghosh,M.(1991) Bayesian prediction in linear models: Applications to small area estimation, *The Anals of Statistics*, 19, 1748-1770

[11] Datta, G. S., Lahiri, and Maiti, T.(1999) Empirical Bayes estimation of median income of four-person by states using time series and cross-sectional data, *Technical Report, Department of Statistics, University of George-Athens*

[12] Dempster, A. P., and Tomberlin, T. J. (1980) The analysis of census undercount from a post-enumeration survey, *in Proceedings of the Conference on Census Undercount*, pp. 88-94

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

[14] Fay, R. E., and Herriot, R.(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

[15] Ghosh, M. and Lahiri, P.(1987) Robust empirical Bayes estimation of means from stratified samples, Journal of the American Statistical Association, 82, 1153-1162

[16] Ghosh, M., Natarajan, K., Kim, D. and Waller, L.A. (1997) Hierarchical Bayes GLM's for the analysis of spatial data: an application to disease mapping. *Tech. Rep.* , University of Florida-Gainsville

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

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

[19] 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

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

[21] Jiang, J. , Lahiri, P.A. and Wan, S.(1998) Jackknifing the mean squared

error of empirical best predictor, *Technical Report, Department of Statistics, Case Western Reserve University*

[22] J. N. K. Rao(1999) Some recent advances in model-based small area estimation, *Survey Methodology*, Vol25, No.2, 175-186

[23] 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

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

[25] MacGibbon, B., and Tomberlin, T. J. (1989) Small area estimates of proportions via empirical Bayes techniques, *Survey Methodology*, 15, 237-252

[26] Malec, D., Sedransk, J., and Tompkins, L. (1993) Bayesian predictive inference for small areas for binary variables in the National Health Interview Survey, *in Case Studies in Bayesian Statistics*, eds. C. Gatsonis, J. S. Hodges, R. E. Kass, and N. D. Singpurwalla, New York:Springer-Verlag, pp. 377-389

[27] Manas, Lahiri, Michael Larsen, and John Reimnitz(1999) Composite estimation of drug prevalences for sub-state areas, *Survey Methodology*, Vol.25, No.1, 81-86

[28] 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

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

[30] Nandram, B., and Sedransk, J.(1993) Bayesian predictive inference for a finite population proportion: Two-Stage Cluster Sampling, *Journal of the Royal Statistical Society, Ser. B*, 55, 399-408

[31] 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

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

[33] Shao,J., and Tu,D. (1995) The Jackknife and the Bootstrap. New York:Springer-Verlag

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

[35] Stroud, T. W. F. (1987) Bayes and empirical Bayes approaches to small area estimation of small area statistics, *in International Symposium on Small Area Statistics*, eds. R. Paltek, J. N. K. Rao, C.E. Saerndal, and M. P. Singh, New York: Wiley, pp. 124-140

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

[37] You, Y. and Rao, J. N. K. (1999) Pseudo Bayes small area estimation using a simple random effects model and sampling weights. *Tech. Rep.*, Statistics Canada

「소지역통계 추정법(1차 연구결과 자료)」의 내용에 관한 문의 또는 의견이 있으시면 다음 연락처를 이용하여 주시기 바랍니다.

연락처 : 「우 302-701」 대전광역시 서구 둔산동 920번지
정부대전청사 통계청 조사관리과
☎(042)481-2088~2089 Fax(042)481-2464
E-mail : kkyoung@nso.go.kr

▣ 발간에 참여한 사람들

공군사관학교 : 이 계오 교수
조사관리과장 : 김 상식
담당사무관 : 김 규영
담 당 직 원 : 변 루나

인쇄 : 협성문화사 ☎(042)627-8893