응용통계연구 (2008)
21(4), 665–674

일반화추정방정식을 활용한 소지역 추정과 실업률패널분석

여인권[1] · 손경진[2] · 김영원[3]

[1]숙명여자대학교 통계학과; [2]숙명여자대학교 통계학과; [3]숙명여자대학교 통계학과

(2008년 5월 접수, 2008년 5월 채택)

요 약

기존의 소지역추정 연구에서는 대부분 특정 시점에서의 관심 모수를 추정하는 문제를 다루어 왔다. 그러나 대부분의 공식통계들은 월, 분기, 또는 년 단위로 반복적으로 얻어지는 패널자료이기 때문에 이를 고려한 추정방법이 필요하다. 이 논문에서는 반복측정 또는 다시점자료 분석에 유용하게 사용되고 있는 일반화추정방정식을 이용한 실증분석을 통해 소지역추정에서 시간종속성을 포함시키는 방안을 알아본다. 실증분석에서는 2005년 1월에서 12월까지의 경상남도 및 울산광역시 월별 경제활동인구조사 자료를 바탕으로 시군구별 실업률과 실업률에 영향을 줄 것으로 생각되는 설명변수의 관계를 일반화선형모형과 일반화추정방정식을 적용하여 분석해 보고 시간종속성을 고려한 것과 하지 않은 것을 비교해 본다.

주요용어: 소지역추정, 실업률, 일반화선형모형, 일반화추정방정식.

1. 서론

1995년 지방자치제도가 시작되면서 지방자치 행정의 정착을 위하여 광역시 또는 도 단위의 통계뿐만 아니라 시군구 등과 같은 소지역(small area) 통계에 대한 요구가 증대되었다. 또한 소득 양극화 등과 관련해 실업 관련 통계가 정치·사회적인 관심사로 대두되면서 시군구별 실업률 통계에 대한 수요가 급증하게 되었다. 그러나 경제활동인구조사를 비롯한 대부분의 정부통계를 생산하기 위한 표본설계는 광역시 또는 도 단위와 같은 대영역의 통계를 생산할 목적으로 설계되기 때문에, 시군구 등과 같은 소지역의 경우 배정되는 표본 조사구수가 극히 적어 신뢰할 수 있는 통계 산출이 어렵다. 또한 전국의 모든 시군구 통계 작성을 위해 새로운 표본조사를 실시하는 것은 비용을 고려할 때 현실적으로 거의 불가능하다. 따라서 기존 표본설계에서 조사된 자료를 가지고 일정 수준의 정도(precision)를 만족하는 시군구 단위의 통계를 생산할 수 있는 소지역 추정기법에 대한 연구가 요구된다.

소지역추정(small area estimation)이란 배정된 표본크기가 작은 소지역이나 성별, 연령, 교육수준, 소득수준 등과 같은 변수의 특성으로 분류된 소영역(small domain)에 대한 통계를 생산하는데 이용되는 추정방법이다. 외국에서는 분석대상에 따라 다양한 형태의 소지역추정 기법들에 대한 심층적인 연구들

본 연구는 숙명여자대학교 2007년도 교내연구비 지원에 의해 수행되었음.
[1]교신저자: (140-742) 서울특별시 용산구 청파동 2가, 숙명여자대학교 이과대학 통계학과, 부교수.
E-mail: inkwon@sookmyung.ac.kr
[2](140-742) 서울특별시 용산구 청파동 2가, 숙명여자대학교 이과대학 통계학과, 석사과정.
E-mail: 1981son@hanmail.net
[3](140-742) 서울특별시 용산구 청파동 2가, 숙명여자대학교 이과대학 통계학과, 교수.
E-mail: ywkim@sookmyung.ac.kr

이 진행되고 있다. 특히, 통계 선진국인 미국, 캐나다, 영국 등의 경우 정부기관과 전문 학자들과의 공동 연구를 통해 소지역추정 기법에 상당한 이론적 결과들을 축적하여 정부 차원의 통계조사인 인구, 소득, 노동력, 농업, 보건통계 등 다양한 분야에서 이를 활용하고 있다.

소지역추정에는 조사된 자료 그 자체만을 이용하는 직접추정법(direct estimation)과 가용할 수 있는 그 지역의 행정자료나 센서스 자료를 보조정보로 이용하여 추정하는 간접추정법(indirect estimation)이 있으며 간접추정법에는 합성추정법(synthetic estimation)과 복합추정법(composite estimation)이 있다. 합성추정법은 인접 유사지역에 있는 소지역은 비슷한 특성을 갖는다는 가정 하에서 주변이나 유사지역의 정보를 이용하여 관심변수에 대한 추정값의 정도를 높이는 방법이다. 직접추정량은 각 소지역의 표본수가 매우 적은 경우 변동성이 커져 신뢰도가 떨어지고, 합성추정량은 해당 소지역과 인접 유사지역의 정보가 동질적이지 못할 경우 편형이 발생한다는 문제가 있다. 이를 보완하기 위하여 직접추정량과 합성추정량의 가중평균을 이용하는 방법이 복합추정법이다. 모형기반추정법(model-based estimation)은 모형 구조가 소지역 간의 복잡한 오차구조를 내포하고 있기 때문에 소지역 간의 변동을 반영하여 소지역 추정의 정확도를 높일 수 있으며, 표본 자료로부터 모형의 유용성이 확인될 수 있다는 장점이 있다. 또한 연속형 자료뿐만 아니라 이진(binary) 또는 범주형 및 시계열 자료와 같은 다양한 자료에도 적용할 수 있다. 특히, 지수족(exponential family)을 따르는 자료에 포괄적으로 적용할 수 있는 일반화선형모형(generalized linear model)과 일반화혼합선형모형(generalized linear mixed model)에 대한 연구가 Ghosh 등 (1998)에 의해 활발하게 이루어지고 있으며 Rao (2003)은 다양한 형태의 소지역 추정방법에 대한 최근까지의 연구결과들을 체계적으로 정리하였다.

우리나라의 경우에는, 기존의 소지역추정 방법들을 우리나라 자료에 적용한 사례분석에 대한 연구가 많다. 김영원과 성나영 (2000)이 도소매업 사업체 조사를 바탕으로 소지역 추정기법의 도입 가능성을 검토한 이후 박종태와 이상은 (2001)은 경제활동인구조사를 토대로 경기도 시군구의 실업자 총계 추정문제를 제한적인 모의실험을 통해 다루었다. 또한 정연수 등 (2003)은 통계청의 협조를 받아 충청북도 시군구 실업자 총계를 추정하기 위해 정규분포를 가정한 소지역추정모형을 개발했으며, 김영원과 최형아 (2004)는 충북지역 대상으로 소지역추정을 위해 로지스틱 모형을 도입한 사례연구결과를 제시한 바 있다.

지금까지 우리나라에서 실시된 소지역추정에 관한 연구들은 대부분 특정 시점에서의 지역이나 영역에 대한 추정에 국한되어 연구되었다. 그러나 정부통계 자료들은 월, 분기, 또는 년 단위로 지속적으로 얻어지는 패널자료이기 때문에 시계열적 요인에 대한 분석을 소지역 추정에 포함시키는 것이 타당할 것으로 생각된다. 소지역추정에서 시계열 성분을 추가한 모형에 대한 연구가 Rao와 Yu (1994), Datta 등 (1999), You 등 (2003) 등에 의해 이루어졌다. 그러나 이들 모형에서는 시계열성분을 AR(1)에 국한하였고 이론 전개에 있어 1차 자기상관을 알고 있다는 가정을 사용하는 등 실제분석에 있어 실용성이 떨어진다. 김재두 등 (2005)은 격자(lattice) 자료분석을 위한 공간시계열모형을 이용하여 전국과 시도별 실업자 수에 대해 시계열분석을 하였으나 모형설정에 있어 근거가 약하고 지역별 모형적합이 개별적으로 이루어져 이론적 도출에 한계가 있었다. 본 논문은 앞서 연구된 논문들과 달리 시계열적 요인을 직접적으로 모형에 추가하여 사용하는 것이 아니라 다시점 또는 경시적 자료(longitudinal data)로 간주하고 일반화추정방정식을 이용하여 소지역 추정을 실시한다. 자료수집의 한계로 2005년 1월부터 12월까지의 월별 경상남도 및 울산광역시 경제활동인구조사 자료만을 사용하여 시군구 실업률 관계식을 유도해 보았다.

2. 일반화추정방정식

일반화선형모형에서는 관심의 대상인 반응변수는 지수족에 속하는 확률분포를 따른다고 가정 하에서 가능도함수(likelihood function)를 근거로 모수에 대한 추론이 이루어진다. 일반화선형모형은 이진자료를 포함한 다양한 형태의 반응변수에 대해 모형화 할 수 있는 장점이 있으나 반응변수의 분포에 대한 정보가 없거나, 이 논문에서 분석할 경제활동인구 자료처럼, 다시점 자료와 같이 자료들 간에 종속성이 존재하는 경우 가능도 함수가 복잡하고 계산이 어렵기 때문에 다른 통계적 방법이 요구된다.

반응변수 Y_i의 기대값이 $E(Y_i) = \mu_i$이고 이 값이 설명변수 $\boldsymbol{x}_i$들에 영향을 받고 미분가능한 연결함수 h에 의해 $h(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$의 관계식이 성립한다고 할 때, Wedderburn (1974)은 일반화선형모형의 추정방정식이 평균과 분산간의 관계에만 의존한다는 사실에 주목하여 이 관계를 이용하여 모수에 대한 추론을 할 수 있게 하는 로그준가능도함수(log quasi-likelihood function)를 다음과 같이 소개하였다.

$$Q(\mu, y) = \int_y^\mu \frac{y-t}{\phi V(t)} dt,$$

여기서 $V(t)$는 분산함수이고, 퍼짐(dispersion)모수 ϕ에 대해, $\mathrm{Var}(Y_i) = \phi V(\mu_i)$가 성립한다. Liang과 Zeger (1986)는 반복측정된 형태를 가지는 다시점 자료의 분석에 준가능도함수를 사용하여 모수를 추정하는 일반화추정방정식(generalized estimating equations: GEE)이라 불리는 기법을 제안하였다.

확률벡터 $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$는, $i = 1, \ldots, m$, i번째 개체를 n_i번 반복 또는 시간에 따라 관측한 반응변수로 이루어져 있고 각각의 반응변수의 주변기대값 $E(Y_{ij}) = \mu_{ij}$가 설명변수 $\boldsymbol{x}_{ij}$에 영향을 받고 그 관계식이 $h(\mu_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}$로 표시된다고 할 때, 일반화추정방정식 방법에서는 모수 $\boldsymbol{\beta}$의 추정값은 다음과 같은 준점수방정식(quasi-score equation)의 해를 계산하여 구한다.

$$\boldsymbol{S}(\boldsymbol{\beta}) = \sum_{i=1}^{m} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)^T \mathrm{Cov}(\boldsymbol{Y}_i)^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}.$$

이 방정식에서 $\partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}$ 부분은 연결함수로부터 유도할 수 있다. 일반화선형모형에서는 자료들이 독립이므로 $\mathrm{Cov}(\boldsymbol{Y}_i)$는 대각행렬이 되고 대각원소인 분산은 평균과 분산의 관계를 이용하여 모형화 할 수 있으나 다시점 자료의 경우 자료들 간의 종속성이 존재하여 단순히 평균과 분산의 관계식만으로 $\mathrm{Cov}(\boldsymbol{Y}_i)$를 표시할 수 없다. 즉, 일반화추정방정식은 확률벡터 $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$의 결합분포에 대한 정보는 필요하지 않지만 로그준가능도함수를 유도할 때 필요했던 연결함수와 평균과 분산의 관계에 추가로 반응변수들 간의 상관관계를 나타내는 상관행렬의 구조가 필요하다. 이때 가정하는 상관행렬을 가상관행렬(working correlation matrix)이라고 한다.

행렬 $\mathbf{A}_i$를 j번째 대각원소가 $V(\mu_{ij})$인 대각행렬이라고 하면, 공분산행렬 $\mathrm{Cov}(\boldsymbol{Y}_i)$는 다음과 같이 쓸 수 있다.

$$\mathrm{Cov}(\boldsymbol{Y}_i) = \phi\mathbf{A}_i^{\frac{1}{2}}\mathbf{R}_i(\boldsymbol{\alpha})\mathbf{A}_i^{\frac{1}{2}} = \phi\mathbf{V}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}),$$

여기서 $\mathbf{R}_i(\boldsymbol{\alpha})$는 $\boldsymbol{Y}_i$의 가상관행렬을 나타내며 $\boldsymbol{\alpha}$는 가상관행렬의 원소를 모형화하는데 사용되는 추정해야 할 미지의 모수를 나타낸다. 가상관행렬은 반복 측정된 자료의 특징을 고려하여 선택할 수 있다. 이 $\boldsymbol{\alpha}$의 형태에 따라 다양한 가상관행렬이 만들어지는데 비상관(uncorrelated) 또는 단위(identity)행렬, 교환가능(exchangeable)행렬, 정상 1차 자기상관(AR(1))행렬, 그리고 비모수적(nonparametric)행렬 등이 일반적으로 사용되고 있다.

- 비상관(uncorrelated) / 단위(identity) 행렬

$$\mathbf{R}_i(\boldsymbol{\alpha}) = \mathbf{I}_{n_i},$$

여기서 $\mathbf{I}_{n_1}$는 $n_i \times n_i$ 단위행렬이다. 이 행렬은 반복 측정에 따른 종속성을 고려하지 않기 때문에 일반적인 일반화선형모형과 같아지며, 잔차를 통해 보다 정확한 상관행렬을 알고자 할 경우 반복과정에서 초기값으로 많이 사용된다.

- 교환가능(exchangeable) 행렬

$$\mathbf{R}_i(\alpha) = \begin{bmatrix} 1 & \alpha & \cdots & \alpha \\ \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \cdots & 1 \end{bmatrix}.$$

모든 상관계수가 같은 형태를 가지며 랜덤효과구조(random effects structure) 또는 복합대칭(compound symmetry)라고도 부른다. 이 행렬은 가족구성원처럼 유사한 부분단위들에 대해 반복측정이 이루어진 경우에 많이 사용된다.

- 정상 1차 자기상관(Stationary AR(1)) 행렬

$$\mathbf{R}_i(\alpha) = \begin{bmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{n_i-1} \\ \alpha & 1 & \alpha & \cdots & \alpha^{n_i-2} \\ \alpha^2 & \alpha & 1 & \cdots & \alpha^{n_i-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha^{n_i-1} & \alpha^{n_i-2} & \alpha^{n_i-3} & \cdots & 1 \end{bmatrix}.$$

이 행렬은 시계열 자료를 설명하는데 많이 사용되고 있다.

- 비모수적(nonparametric) 구조

$$\mathbf{R}_i(\boldsymbol{\alpha}) = \begin{bmatrix} 1 & \alpha_{12} & \alpha_{13} & \cdots & \alpha_{1n_i} \\ \alpha_{12} & 1 & \alpha_{23} & \cdots & \alpha_{2n_i} \\ \alpha_{13} & \alpha_{23} & 1 & \cdots & \alpha_{3n_i} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha_{1n_i} & \alpha_{2n_i} & \alpha_{3n_i} & \cdots & 1 \end{bmatrix}.$$

이 행렬은 α_{ij}들에 대해 안정적인 추정값을 얻을 수 있을 정도로 자료가 충분한 경우에 사용할 수 있다.

각각의 행렬에 대한 모수 α와 ϕ에 대한 추정값은 Liang과 Zeger (1986)에 제시되어 있다.

선택된 가상관행렬 $\mathbf{R}(\alpha)$에 대해, 일반화추정방정식 방법에 의한 추정값 $\hat{\boldsymbol{\beta}}$와 $\hat{\alpha}$는 다음의 추정방정식의 해이다.

$$\boldsymbol{S}(\boldsymbol{\beta}, \alpha) = \sum_{i=1}^{m} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)^T \mathbf{V}_i(\boldsymbol{\beta}, \alpha)^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}.$$

일반화추정방정식으로부터 구한 모수의 추정량은 연결함수가 정확할 때 일치추정량이 되고 근사적으로 정규분포를 따르게 된다.

3. 경제활동 패널자료를 이용한 실업률 추정

현재 통계청에서는 취업과 실업 등과 같은 경제활동인구조사를 매월 표본가구 내에 거주하는 만 15세 이상인 사람들을 대상으로 실시하고 있다. 최근에 시군구 지역단위의 실업 관련 통계에 대한 관심도가

표 3.1. 월별 실업률 추이(%)

지역	월	2월	3월	4월	5월	6월	7월	8월	9월	10월	11월	12월
A	4.3	1.9	2.5	3.7	2.4	2.5	1.3	2.7	2.6	1.9	2.2	2.2
B	2.0	1.9	0.0	1.6	1.6	3.1	0.0	0.0	0.0	3.0	2.8	2.8
C	2.8	2.9	1.4	2.8	2.7	4.2	2.8	2.7	1.3	1.3	0.0	0.0
D	5.6	3.7	2.5	2.8	2.4	2.9	2.8	2.0	2.0	2.4	3.5	3.5
E	4.8	4.5	5.3	4.6	4.4	4.0	4.9	4.6	3.8	3.9	3.7	3.7
F	0.0	1.8	3.6	1.8	0.0	1.8	0.0	0.0	0.0	0.0	0.0	0.0
G	1.8	3.6	4.0	3.3	2.7	3.2	2.1	1.8	2.2	1.1	1.5	1.5
H	4.4	3.6	4.1	2.8	3.4	2.9	2.5	2.9	2.5	2.1	4.0	4.0
I	3.6	4.0	3.2	1.6	0.0	4.4	5.9	5.2	3.0	3.0	3.7	3.7
J	3.2	3.6	3.7	3.9	3.5	1.8	1.3	2.2	1.8	1.3	2.7	2.7
K	3.2	2.2	3.3	2.9	3.8	2.8	2.6	3.7	4.4	1.8	1.4	1.4
L	5.0	2.5	3.2	1.8	0.6	0.7	2.0	4.0	2.0	3.5	2.9	2.9
M	1.7	2.8	3.1	2.7	2.3	3.0	5.0	3.7	4.3	3.3	2.1	2.1
N	2.9	0.0	5.0	5.1	2.5	0.0	0.0	2.6	2.9	0.0	1.5	1.5
O	6.3	4.8	3.2	2.4	4.5	3.6	3.6	4.2	3.6	3.3	3.3	3.3
P	1.6	1.2	0.8	1.2	0.4	0.0	1.9	1.6	2.0	2.3	2.5	2.5
Q	1.9	2.0	1.3	2.0	3.4	6.9	3.2	1.9	1.0	4.8	2.6	2.6
R	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	1.9	2.4	2.4
S	3.6	4.5	3.6	3.3	2.9	3.7	4.1	3.2	3.0	4.3	2.7	2.7
T	0.9	2.5	2.4	4.2	4.2	1.8	3.7	1.0	3.6	0.0	1.0	1.0
U	1.7	1.7	0.0	0.0	1.6	1.6	1.6	1.8	1.7	0.0	0.8	0.8
V	3.6	3.2	3.3	4.0	3.7	0.0	0.0	0.0	3.6	0.0	1.8	1.8
W	3.4	0.0	0.0	1.3	1.3	1.4	0.0	0.0	1.3	1.3	1.6	1.6

점점 높아지고 있지만, 현재 우리나라에서 실시되는 경제활동인구조사의 표본설계는 관심영역이 광역시 또는 도와 같은 대영역이므로 시군구와 같은 소지역에 해당되는 각 행정자치구역들은 표본설계에 고려되지 않고 있다. 따라서 현재 활용되고 있는 대영역 기반의 표본설계를 이용하여 소지역 통계를 직접 생산할 경우, 시군구 지역에 대해서는 표본조사구의 수가 하나 내지 두 개 정도로 너무 작게 배정되는 경우가 다수 발생하기 때문에 신뢰할 만한 소지역 단위의 통계생산은 어렵게 된다.

본 논문에서는 2005년 1월부터 12월까지 경제활동인구조사 자료에서 경상남도 및 울산광역시를 연구대상으로 삼았다. 경상남도는 10개의 시와 10개의 군으로 이루어져 있으며 울산광역시는 4개의 구와 한 개의 군으로 구성되어 있다. 경제활동인구조사에서는 표본조사구별로 조사대상의 성별, 학력수준, 연령, 경제활동여부, 실업자여부에 대한 정보가 얻어진다. 이 실증분석에서는 지역별로 시간의 흐름에 따른 실업률의 추이를 살펴보고 신기일과 이상은 (2003)이 소지역에 근접한 지역의 정보를 이용한 모형을 사용한 것에서 착안하여 특정 지역의 실업률이 인접한 지역들의 실업률과 관계가 있는지 살펴보았다. 더불어 지역별 실업률과 실업률에 영향을 줄 것으로 예상되는 설명변수의 관계를 일반화선형모형과 일반화추정방정식을 이용하여 설명변수들이 실업률에 어떠한 영향을 끼치는지 살펴보았다.

3.1. 시군구별 실업률 추이

본 논문에서 사용하는 '실업률'은 고용통계에서 정의한 실업자가 경제활동인구에서 차지하는 비율을 의미한다. 표 3.1는 2005년 1월 부터 12월까지 시군구별 실업률 추이를 정리한 것이다. 조사기간 동안

표 **3.2.** 경계를 기준으로 한 이웃정보

지역번호	지역명	주변지역	지역번호	지역명	주변지역
01	거제시	21	02	거창군	12 24 25
03	고성군	08 11 17 21	04	김해시	09 13 18 20
05	남구	07 10 14 16	06	남해군	22
07	동구	05 10	08	마산시	03 17 20 23
09	밀양시	04 13 14 19 20	10	북구	05 07 14 16
11	사천시	03 17 22	12	산청군	02 15 17 22 24 25
13	양산시	04 09 14	14	울주군	05 09 10 13 16
15	의령군	12 17 19 23 25	16	중구	05 10 14
17	진주시	03 08 11 12 15 22 23	18	진해시	04 20
19	창녕시	09 15 20 23 25	20	창원시	04 08 09 18 19 23
21	통영시	01 03	22	하동군	06 11 12 17 24
23	함안군	08 15 17 19 20	24	함양군	02 12 22
25	합천군	02 12 15 19			

Q지역의 6월 실업률이 6.9%로 가장 높은 것으로 나타났으며 두 개의 군 지역은 조사기간 동안 조사대상 경제활동인구 중 실업자가 한 명도 없는 것으로 나타나 표에 표시하지 않았다.

월별 실업률과 경제활동비율과의 관계를 분석한 결과, 대체적으로 삼각형 형태의 분포를 가지는 것으로 확인되었는데 이것은 경제활동비율이 상대적으로 낮거나 높게 나타날 경우 실업률이 대체로 낮게 나타나는 것을 의미한다. 경제활동비율이 60%정도일 경우 대체로 실업률이 높게 나타나는 형태를 보였다.

3.2. 공간상관관계

실업여부에 성별이나 학력수준 등의 변수가 영향을 줄 수 있으며 또한 지역의 위치정보가 해당지역의 실업률 추정에 영향을 줄 수도 있다. 본 논문에서는 분석의 정확성을 높이기 위해서, 해당 시군구와 경계를 가지고 있는 시군구를 주변지역으로 정의하고, 특정 시군구의 지역적 특성을 반영하기 위해 주변 시군구의 평균실업률을 모형에 반영하였다. 조사구가 거제시나 남해군과 같이 섬에 있는 경우에는 거제대교나 남해대교와 같이 다리로 연결된 지역을 주변지역으로 정하였다. 표 3.2는 각 조사지역에 대한 주변지역을 나타낸 것이다.

해당지역의 실업률과 주변지역의 평균실업률 간의 관계에 있어서도 해당지역의 실업률과 경제활동비율과의 관계와 대체로 유사한 양상을 나타내는 것으로 나타났다. 즉 주변지역의 평균실업률이 상대적으로 낮거나 높게 나타날 경우 대체로 실업률이 낮게 나타나는 것으로 보였다.

3.3. 일반화선형모형을 활용한 분석

일반화선형모형은 자료가 서로 독립이라는 가정하에서 분석이 가능하다. 그러나 동일 조사구 또는 지역에서 얻어진 월별 자료는 서로 독립이라고 할 수 없으므로 월별로 분석이 수행되어야 한다. 이런 모형에서 지역특성을 반영하기 위해 해당지역의 시군구별 조사대상 경제활동인구에서의 남성비율, 평균학력수준, 평균연령, 행정구역형태, 주변지역의 평균실업률을 설명변수로 사용했다. 여기서 평균학력수준은 중졸미만, 중졸, 고졸, 초대졸, 대졸, 대학원이상으로 구분하였으며 행정구역형태는 광역시, 시, 군에 따라 두 개의 가변수로 표시하였다. 반응변수는 경제활동인구 중 실업자의 수로 이항분포를 따른다고 가정하고 이에 따라 연결함수는 이진자료분석에서 일반적으로 많이 사용되고 있는 logit를 적용하였다.

표 **3.3.** 10%유의수준에서 유의한 월별 회귀계수추정값

월	상수항	시	군	남성비율	평균연령	평균학력	주변실업률
1	−13.714	-0.466^b	-1.221^a	·	0.136^c	1.830^b	·
2	−5.980	·	·	·	·	0.293^a	·
3	−5.088	-0.630^b	-0.627^c	·	·	0.974^b	-23.408^c
4	−0.862	·	·	·	-0.062^b	·	·
5	−3.191	-0.527^a	-0.893^a	·	·	·	·
6	−4.318	·	·	·	·	·	27.226^b
7	−0.360	·	·	·	-0.074^a	·	·
8	−9.339	·	·	10.682^b	·	·	25.943^b
9	−6.733	·	·	6.640^c	·	·	·
10	−6.592	·	·	·	·	1.068^a	·
11	−8.342	·	·	·	·	1.352^a	33.312^a
12	−9.387	-0.380^b	-1.202^a	-12.057^b	0.123^b	2.412^a	·

표 3.3은 월별자료를 독립적인 12개의 데이타세트로 나누고 월별로 로지스틱회귀모형에 적합 시켜 구한 회귀계수의 추정값이다. 사용된 설명변수간의 상관관계가 높아 다중공선성과 같은 문제가 발생하여 모든 변수를 모형에 포함시킨 경우 대부분의 분석에서 유의한 회귀계수가 없는 것으로 나타났다. 이 논문에서는 이를 쉽게 해결하기 위해 후진소거법을 이용하여 10% 유의수준에서 의미 있는 유의한 회귀계수의 변수를 선택하였다. 표에서 a는 1%, b는 5%, c는 10% 유의수준에서 설명력이 있는 계수를 의미한다. 예를 들어 3월 자료의 경우, 시와 군의 가변수, 평균학력수준, 주변지역의 평균실업률이 유의한 설명변수인 것으로 분석되었다.

위의 표 3.3에서 볼 수 있는 것과 같이 월별로 유의한 설명변수가 다르게 선택되었으며 그 차이가 큰 것으로 나타났다. 광역시의 경우 시와 군보다 상대적으로 실업률이 높고 평균학력이 높을수록 실업률이 높다는 부분적인 해석이 가능할 것으로 생각되나 일반화선형모형을 이용한 월별 분석에서는 전체적으로 일관성 있는 해석이 어렵다. 월별분석에 사용된 자료가 25개로 많지 않은 것이 안정적인 모형 선택이 이루어지지 않은 이유가 될 수도 있다. 따라서 소지역추정모형 개발에 있어 월별자료를 독립적으로 처리하는 일반화선형모형 방법은 적절하지 않은 것으로 판단된다.

3.4. 일반화추정방정식을 활용한 분석

동일 지역에서의 월별 시계열 변동과 시군구별 특성을 모형에 동시에 반영하고 적용하기 위해서는 시계열 또는 반복측정에 따른 종속성을 종합적으로 반영한 일반화추정방정식을 소지역추정모형에 활용하는 것이 효과적일 수 있다. 일반화추정방정식을 활용한 분석에서 사용된 변수는 일반화선형모형에서 사용된 변수와 동일하다. 표 3.4는 연결함수를 logit, probit, cloglog로 가정하고 각각의 연결함수에 대해 가상관행렬을 교환가능과 AR(1)일 때를 가정하여 적합하였을 때 모수추정값을 정리한 것이다. 여기에서도 설명변수들간의 상관관계가 높아 10% 유의수준 후진소거법에 의해 설명변수를 선택하였다. 표 3.4에서 a는 1%, b는 5%, c는 10% 유의수준에서 설명력이 있는 계수를 의미한다.

표 3.4의 결과를 보면 일반화추정방정식 모형에서 logit, probit, cloglog 등의 연결함수를 사용함에 따라 나타난 변수선택 상의 차이는 없는 것으로 나타났다. 예를 들어, 교환가능 가상관행렬을 가정하는 경우 시군구별 실업률을 설명함에 있어 어떤 연결함수를 사용하던지 군, 평균연령, 주변평균실업률 등의 설명변수가 유의한 것으로 나타났다. 평균교육수준과 남성비율의 상관관계는 0.703, 평균교육수준과 평균연

표 **3.4.** 연결함수와 가상관행렬에 따른 모수추정값

연결함수	logit		probit		cloglog	
상관행렬	교환가능	AR(1)	교환가능	AR(1)	교환가능	AR(1)
$\hat{\alpha}$	0.314	0.512	0.310	0.509	0.314	0.512
상수항	-1.921	-4.902	-1.256	-2.474	-1.952	-4.900
시	·	-0.183^b	·	-0.078^b	·	-0.180^b
군	-0.263^b	-0.434^a	-0.109^b	-0.181^a	-0.260^b	-0.428^a
남성비율	·	·	·	·	·	·
평균연령	-0.044^a	·	-0.018^a	·	-0.044^a	·
평균학력	·	0.553^a	·	0.231^a	·	0.546^a
주변실업률	10.159^a	·	4.313^a	·	10.011^a	·

령의 상관관계는 -0.946, 평균연령과 남성비율의 상관관계는 -0.734로 이들 변수 간에 밀접한 관계가 있는 것을 확인하였다. 이것은 변수선택에서 있어 가장 설명력이 높은 변수 하나에 의해 실업률에 대한 다른 두 변수의 영향이 어느 정도 설명될 수 있다는 것을 의미하며 표 3.4에서 보는 것과 같이 셋 중에 한 개의 변수만이 유의한 것으로 나타난 이유가 된다.

동일 지역에서 월별 반복측정에 따른 종속성은 표 3.4의 가상관행렬 계수에 해당하는 $\hat{\alpha}$로 설명될 수 있다. 월간 상관계수가 모두 동일하다는 교환가능행렬 가정 하에서는 상관계수는 logit, probit, cloglog 연결함수에 대해 각각 $\hat{\alpha} = 0.314, 0.310, 0.314$로 추정되었으며 행정구역이 군인 경우, 평균연령, 주변 실업률이 유의한 것으로 나타났다. 행정구역이 군인 지역이 광역시나 시보다 실업률이 낮은 것으로 나타났고 평균연령이 높을수록 실업률은 낮아지고 주변지역의 실업률이 높을수록 실업률이 높은 것으로 나타났다. 가상관행렬을 AR(1)으로 가정할 경우에는 각각 $\hat{\alpha} = 0.512, 0.509, 0.512$로 추정되었고 모든 연결함수에 대해 행정구역과 평균학력수준이 유의한 설명변수인 것으로 확인되었다. 광역시에 비해 시와 군이 실업률이 낮은 것으로 나타났으며 특히 군인 경우 더욱 낮은 것으로 분석되었다. 평균학력이 높은 지역일수록 실업률이 높았는데 이것은 평균교육수준과 상관관계가 -1에 가까운 평균연령으로 해석된 교환가능 상관행렬 결과와 큰 차이가 없다고 볼 수 있으나 주변지역의 실업률에 영향을 받지 않는다는 것이 가장 큰 차이점이라고 할 수 있다.

일반화추정방정식은 준가능도함수를 바탕으로 구해지기 때문에 가능도함수를 기반으로 한 AIC나 SBC와 같은 가능도함수를 근거로 한 모형비교는 할 수 없다. 가상관행렬의 가정에 따라서 결과가 조금씩 다르게 나타났으나 평균교육수준 또는 평균연령, 행정구역(시군구), 주변지역의 실업률이 각 지역의 실업률에 영향을 주는 것으로 보인다.

4. 결론

본 논문에서는 2005년 1월부터 12월까지 경상남도 및 울산광역시의 25개 시군구의 경제활동인구조사 자료를 일반화선형모형과 일반화추정방정식을 활용하여 시군구 소지역 실업률에 어떤 설명변수들이 어떻게 영향을 끼치는지 살펴보았다. 각각의 월별 자료에 대한 일반화선형모형 분석에서는 일관성 있는 변수선택이 어려운 것을 확인하였다. 또한 교환가능 가상관행렬과 AR(1) 가상관행렬의 추정값은 지역 내에서의 월별 실업률 간에 상관관계가 높다는 것을 확인하였다. 이것은 일반화선형모형이나 비상관행렬을 가정한 일반화추정방정식이 현재 자료에 적절하지 않다는 것을 의미한다. 분석에서 언급된 연결함수에 대해 동일한 설명변수가 선택되었는데 해당지역의 실업률은 학력수준 또는 연령, 행정구역, 주변지역의 실업률에 영향을 받는 것으로 나타났다.

참고문헌

김재두, 신기일, 이상은 (2005). 공간 시계열 모형을 이용한 소지역 추정, <응용통계연구>, **18**, 627–637.

김영원, 성나영 (2000). 소지역 통계 생산을 위한 추정 방법, *Journal of the Korean Data & Information Science Society*, **11**, 111–126.

김영원, 최형아 (2004). Small area estimation techniques based on logistic model to estimate unemployment rate, *The Korean Communications in Statistics*, **11**, 583–595.

박종태, 이상은 (2001). 소지역 추정법에 관한 비교연구, *Journal of the Korean Data & Information Science Society*, **12**, 47–55.

신기일, 이상은 (2003). Model-data based small area estimation, *The Korean Communications in Statistics*, **10**, 637–645.

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

Datta, G. S., Lahiri, P., Maiti, T. and Lu, K. L. (1999). Hierarchical Bayes estimation of unemployment rates for the States of the U.S., *Journal of the American Statistical Association*, **94**, 1074–1082.

Ghosh, M., Natarajan, K., Stroud, T. W. F. and Carlin, B. P. (1998). Generalized linear models for small-area estimation, *Journal of the American Statistical Association*, **93**, 273–282.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika*, **73**, 13–22.

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

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

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear model and the Gauss-Newton method, *Biometrika*, **61**, 439–447.

You, Y., Rao, J. N. K. and Gambino, J. (2003). Model-based unemployment rate estimation for the Canadian labour force survey: A hierarchical Bayes approach, *Survey Methodology*, **29**, 25–32.

Small Area Estimation via Generalized Estimating Equations and the Panel Analysis of Unemployment Rates

In-Kwon Yeo[1] · Kyoungjin Son[2] · Youngwon Kim[3]

[1]Dept. of Statistics, Sookmyung Women's University;
[2]Dept. of Statistics, Sookmyung Women's University;
[3]Dept. of Statistics, Sookmyung Women's University

(Received May 2008; accepted May 2008)

Abstract

Most of existing studies about the small area estimation deal with the estimation of parameters based on cross-sectional data. However, since many official statistics are repeatedly collected at a regular interval of time, for instance, monthly, quarterly, or yearly, we need an alternative model which can handle characteristics of these kinds of data. In this paper, we investigate the generalized estimating equation which can model time-dependency among response variables and is useful to analyze repeated measurement or longitudinal data. We compare with the generalized linear model and the generalized estimating equation through the estimation of unemployment rates of 25 areas in Gyeongsangnam-do and Ulsan. The data consist of the status of employment and some covariates from January to December 2005.

Keywords: Generalized estimating equations, generalized linear models, small area estimation, unemployment rates.

This research was supported by the Sookmyung Women's University Research Grants 2007.
[1]Corresponding author: Associate Professor, Dept. of Statistics, Sookmyung Women's University, Chungpa-dong 2-ga, Yongsan-gu, Seoul 140-742, Korea. E-mail: inkwon@sm.ac.kr
[2]Graduate student, Dept. of Statistics, Sookmyung Women's University, Chungpa-dong 2-ga, Yongsan-gu, Seoul 140-742, Korea. E-mail: 1981son@hanmail.net
[3]Professor, Dept. of Statistics, Sookmyung Women's University, Chungpa-dong 2-ga, Yongsan-gu, Seoul 140-742, Korea. E-mail: ywkim@sm.ac.kr