Journal of the Korean Data & Information Science Society
2013, **24**(6), 1199–1210

http://dx.doi.org/10.7465/jkdi.2013.24.6.1199
한국데이터정보과학회지

소지역 실업률의 패널추정을 위한 일반화커널추정방정식[†]

심주용[1] · 김영원[2] · 황창하[3]

[1]인제대학교 데이터정보학과 · [2]숙명여자대학교 통계학과 · [3]단국대학교 응용통계학과

접수 2013년 6월 29일, 수정 2013년 8월 4일, 게재확정 2013년 8월 23일

요 약

오늘날 높은 실업률은 대부분의 국가에서 중요한 문제 중의 하나이다. 한편 소지역의 노동 관련 통계에 대한 요구가 지난 몇년간 급속도로 증가하였다. 그러나 대부분의 공식통계를 생산하기 위한 표본설계는 대영역의 통계를 생산할 목적으로 설계되기 때문에 소지역의 경우 배정되는 표본조사단위 수가 극히 적어 신뢰성 있는 통계 산출이 어렵다. 그리고 소지역 추정에 대한 대부분의 기존 연구들은 특정 시점에서의 추정에 국한 되어 왔다. 그러나 대부분의 공식통계들은 월, 분기 또는 연 단위로 측정되는 패널자료이기 때문에 이를 고려한 추정방법이 필요하다. 본 논문에서는 패널자료의 분석을 위해 유용하게 사용되고 있는 일반화추정방정식의 비모수적 버전인 일반화커널추정방정식을 도출하여 조사시점을 고려한 소지역 실업률의 추정에 활용하는 방안을 제안한다. 모의실험을 통하여 일반화커널추정방정식 방법, 일반화추정방정식 방법 및 일반화선형모형과 비교한다. 그리고 2005년 1월부터 12월까지 경상남도 및 울산광역시의 25개 시군구의 경제활동인구조사의 패널자료에 위에서 언급한 세 가지 방법을 적용하여 해당 소지역의 월별 실업률을 추정한다.

주요용어: 소지역 추정, 일반화선형모형, 일반화추정방정식, 일반화커널추정방정식, 커널기법, 패널자료, 패널추정.

1. 서론

통계청에 의해 매월 15일이 포함된 1주일 동안 전국 약 32,000 표본가구에 대해 실시되는 경제활동인구조사는 취업, 실업, 노동력 등과 같은 국민의 경제활동의 특성을 조사함으로써 거시경제 분석과 인력자원의 개발정책 수립에 필요한 기초자료를 제공한다는 측면에서 오늘날 매우 중요하다. 특히 실업률은 국가 경제와 관련된 주요지표이기 때문에 통계청에서는 특별시, 광역시 또는 도와 같은 대영역 (large domain) 단위의 실업률을 발표하고 있다. 그러나 지방자치제도가 정착되면서 시군구 등과 같은 소지역 (small area) 통계에 대한 관심과 요구가 증가하고 있다. 한편 국내외적으로 소지역 통계에 관한 관심이 높아지고 있으며 이와 관련하여 소지역 추정 (small area estimation)에 관한 많은 연구가 활발하게 진행되고 있다. 특히 미국, 캐나다, 이탈리아, 영국 등과 같은 통계 선진국에서는 분석대상에 따라 다양한 형태의 소지역 추정 관련 통계적 방법론에 대해 정부기관과 전문 학자들과의 공동연구들이 진행되고 있다. Pfeffermann (2013)과 Rao (2003)는 소지역 추정에 대한 최근까지의 많은 연구결과를 체

[†] 이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임 (NRF-2011-0021125).
[1] (621-749) 경남 김해시 인제로 197, 인제대학교 데이터정보학과, 겸임교수.
[2] (140-742) 서울특별시 용산구 청파로47길 100, 숙명자대학교 통계학과, 교수.
[3] 교신저자: (448-701) 경기도 용인시 수지구 죽전로 152, 단국대학교 응용통계학과, 교수.
E-mail: chwang@dankook.ac.kr

계적으로 정리하여 소개하였다. 특히, 소지역 실업률의 추정과 관련된 연구결과들을 소개하면 Chung 등 (2003), Datta 등 (1999), Khoshgooyanfard와 Monazzah (2006), Kim과 Choi (2004), Pereira 등 (2013), Ugarte 등 (2009), Yeo 등 (2008), You 등 (2003) 등이 있다.

소지역 추정이란 용어는 다소 혼동을 주는 용어이다. 왜냐하면 소지역 추정에서 문제가 되는 것은 해당 소지역의 표본크기이지 소지역 자체의 크기가 아니기 때문이다. 또한 소지역이란 반드시 지리학적 지역만을 의미하는 것이 아니고 사회인구학적 집단 또는 산업형태에 따른 집단 등과 같은 다양한 영역(domain)을 나타내기도 한다. 경제활동인구조사를 포함한 대부분의 공식통계를 생산하기 위한 표본설계는 시·도를 중심으로 한 대영역의 통계를 생산할 목적으로 설계되기 때문에 소지역 추정에서는 계획에 없던 지역을 추정해야 하는 어려움이 있으며 가장 일반적으로 발생하는 어려움은 해당 소지역에 배정되는 표본조사단위수가 극히 적어 신뢰할 수 있는 통계를 산출하지 못한다는 것이다. 이러한 관점에서 대영역 기반의 표본설계 하에서 해당 소지역에 대한 행정자료나 센서스 자료를 보조정보로 이용하여 추정의 정확도를 높이는 소지역 추정법들의 필요성이 제기되고 있다.

일반적으로 소지역 추정법은 크게 직접추정법 (direct estimation), 간접추정법 (indirect estimation)과 모형기반추정법 (model-based estimation)으로 나눌 수 있다. 직접추정법은 조사된 자료 그 자체 만을 이용하는 추정법이며 간접추정법은 해당 소지역의 행정자료 또는 센서스 자료와 인근 소지역에 관한 통계정보를 보조정보로 이용하는 추정법이다. 그리고 간접추정법에는 합성추정법 (synthetic estimation)과 복합추정법 (composite estimation)이 있다. 합성추정법은 추정하고자하는 소지역과 인근 소지역 또는 특성이 유사한 소지역들의 정보를 이용하여 관심변수의 추정값의 정도를 높이고자 하는 추정법이다. 직접추정량은 편향이 없는 추정량이지만 해당 소지역에 배정된 표본의 크기가 작은 경우에는 추정량의 분산이 커져서 신뢰성이 떨어지게 된다. 한편 합성추정량은 해당 소지역과 인근 유사지역의 정보가 동질적이지 못할 경우 편향이 발생하는 문제점이 있다. 이런 문제점을 보완하기 위해 두 추정량의 가중평균을 사용하는 방법이 복합추정법이다. 모형기반추정법도 센서스 자료, 지역의 행정자료, 인근 소지역에 관한 통계정보 또는 이전의 연구결과 등과 같은 추가적인 정보들을 보조정보로 이용하기 때문에 보다 정밀한 추정이 가능하다. 모형기반추정법에서는 보조정보의 양이 많을수록 정확한 소지역 추정이 가능하므로 보조정보의 양을 증가시키는 방법이 연구되고 있다. 또한 소지역 간의 변동성을 오차구조에 반영하여 소지역 추정의 정확도를 높일 수 있다. 더욱이 이진 (binary), 범주형 또는 시계열 자료와 같은 다양한 자료에도 적용할 수 있을 뿐만 아니라 지수족 (exponential family)에 속하는 분포를 따르는 자료에 적용할 수 있는 일반화선형모형 (generalized linear model; GLM)과 일반화혼합선형모형 (generalized linear mixed model; GLMM)에 대한 연구가 Ghosh 등 (1998), Marker (1999), Noble 등 (2002) 등에 의해 활발하게 이루어지고 있다. 국내에서는 Yeo 등 (2008)이 패널자료의 분석을 위해 유용하게 사용되고 있는 일반화추정방정식 (generalized estimating equation; GEE)을 소지역 추정에 활용하였다.

대부분의 소지역 추정법들은 모수적 모형에 기초를 두고 있는데 최근 비모수적 방법이 소지역 추정에 적용되기 시작하였다. Jeong과 Shin (2012), Opsomer 등 (2008), Salvati 등 (2010), Shim과 Hwang (2012)이 비모수적 방법을 이용한 소지역 추정에 관하여 연구결과를 발표하였다. 지금까지의 비모수적 소지역 추정법들은 특정 시점에서의 추정에 국한되어 연구되었다. 본 논문에서는 통계적 학습이론에서 많이 활용되고 있는 커널기법을 GEE에 적용하여 비모수적 버전인 일반화커널추정방정식 (generalized kernel estimating equation; GKEE)을 도출하고 조사시점을 고려한 소지역 실업률 추정에 활용하는 방안을 제안하며 그 유효성을 살펴보고자 한다. 본 논문은 다음과 같이 구성되었다. 2절에서 GEE 방법을 간단히 설명하고, 3절에서 커널기법을 적용하여 GEE의 비모수적 버전인 GKEE를 유도하고 조사시점을 고려한 소지역 실업률 추정에 활용하는 방법을 살펴본다. 4절에서 실험연구를 통하여 제안된 방법의 우수성을 고찰하며, 5절에서 결론을 맺는다.

2. 일반화추정방정식

본 절에서는 동일한 개체를 대상으로 연속적으로 조사를 수행하여 동일 개체로부터 얻은 관측자료들 간에 상관관계가 존재하는 일반적인 패널자료에 적용할 수 있는 GEE 방법에 대해 설명한다. 서술방법과 표현을 위해 Yeo 등 (2008)을 참고하였음을 먼저 밝혀둔다.

Nelder와 Wedderburn (1972)이 제안한 GLM에서는 반응변수 y_i가 지수족에 속하는 확률분포를 따르고 기대값 $E(y_i) = \mu_i$가 연결함수 (link function) $g(\cdot)$를 통하여 설명변수들과 선형관계를 가진다는 가정하에서, 즉 $g(\mu_i) = \boldsymbol{x}_i^t\boldsymbol{\beta}$가 성립한다는 가정하에서 가능도함수 (likelihood function)를 근거로 모수에 대한 추론이 이루어진다. GLM은 이진형 (binary), 계수형 반응변수를 포함한 다양한 형태의 반응변수를 모형화 할 수 있는 장점이 있지만 반응변수의 확률분포에 대한 정보가 없어 설명변수들과 반응변수간의 연결함수를 모르거나 본 논문에서 분석할 경제활동인구조사의 패널자료처럼 자료들 간에 상관관계가 존재하는 경우 가능도함수가 복잡하고 계산이 어렵기 때문에 GEE가 제안되었다. Wedderburn (1974)은 GLM의 추정방정식이 평균과 분산간의 관계에만 의존한다는 사실에 주목하여 설명변수들과 반응변수간의 연결함수를 모를 때도 모수에 대한 추론을 가능하게 하는 준가능도함수 (quasi-likelihood function)를 다음과 같이 소개하였다.

$$Q(\mu, y) = \int_y^\mu \frac{y-t}{\phi V(t)} dt, \tag{2.1}$$

여기서 $V(t)$는 분산함수이고 산포모수 (dispersion parameter) ϕ에 대해 $Var(y_i) = \phi V(\mu_i)$가 성립한다. 식 (2.1)의 준가능도함수는 로그가능도함수와 비슷한 성질을 갖는다. 그리고 Liang과 Zeger (1986)는 반복측정 시계열 자료인 패널자료의 분석에 준가능도함수를 사용하여 모수를 추정하는 GEE를 제안하였으며 자료들 간에 가정된 상관관계 모형이 실제와 다른 경우에도 GEE에 의해 추정된 결과는 일치성을 만족하고 추정량의 분포가 점근적으로 정규분포를 따르는 것을 보였다.

이제 패널자료의 분석을 위한 GEE를 좀더 구체적으로 살펴보자. 먼저 y_{ij}는 i번째 개체를 시간에 따라 반복측정할 때 j번째 시점에서 관측된 반응변수를 나타낸다. 이때 $i = 1, \cdots, m$, $j = 1, \cdots, t_i$. 그리고 $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{it_i})^t$는 i번째 개체에 대한 반응변수들의 벡터를, $\boldsymbol{\mu}_i = (\mu_{i1}, \cdots, \mu_{it_i})^t$는 대응되는 평균벡터를, $\boldsymbol{x}_{ij} = (x_{ij1}, \cdots, x_{ijp})^t$는 i번째 개체에 대해 j번째 시점에서 관측된 설명변수들의 벡터를 나타낸다. 그러면 각 주변 반응변수의 기대값 $E(y_{ij}) = \mu_{ij}$와 설명변수벡터 $\boldsymbol{x}_{ij}$간의 관계식이 $g(\mu_{ij}) = \boldsymbol{x}_{ij}^t\boldsymbol{\beta}$일 때 GEE 방법은 다음의 준점수방정식 (quasi-score equation)의 해를 구하여 회귀계수벡터 $\boldsymbol{\beta}$의 추정값을 얻는다.

$$\boldsymbol{S}(\boldsymbol{\beta}) = \sum_{i=1}^{m} \frac{\partial \boldsymbol{\mu}_i^t}{\partial \boldsymbol{\beta}} Cov(\boldsymbol{y}_i)^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}. \tag{2.2}$$

식 (2.2)의 $\partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}$는 연결함수를 이용하여 쉽게 구할 수 있다. GLM에서는 자료들이 독립이므로 $Cov(\boldsymbol{y}_i)$는 대각행렬이 되고 대각원소인 분산은 평균과 분산의 관계를 이용하여 구할 수 있다. 그러나 패널자료의 경우 자료들 간에 상관관계가 있어 단순히 평균과 분산의 관계식만으로 $Cov(\boldsymbol{y}_i)$를 표현할 수는 없다. 따라서 GEE는 반응변수벡터 $\boldsymbol{y}_i$의 결합분포에 대한 정보는 필요하지 않지만 준가능도함수를 유도할 때 필요한 평균과 분산의 관계에 추가로 반응변수들 간의 상관관계를 나타내는 상관행렬의 구조가 필요하다. 이때 가정하는 상관행렬을 가상관행렬 (working correlation matrix)이라고 한다.

행렬 $R_i(\boldsymbol{\rho})$를 $t_i \times t_i$ 가상관행렬이라고 할 때 공분산행렬 $Cov(\boldsymbol{y}_i)$는 다음과 같이 표현될 수 있다.

$$Cov(\boldsymbol{y}_i) = \phi A_i^{\frac{1}{2}} R_i(\boldsymbol{\rho}) A_i^{\frac{1}{2}} = \phi V_i(\boldsymbol{\beta}, \boldsymbol{\rho}), \tag{2.3}$$

여기서 A_i는 j번째 대각원소가 $V(\mu_{ij})$인 $t_i \times t_i$ 대각행렬을 나타내고 $\boldsymbol{\rho}$는 가상관행렬을 완전히 모형화하는데 사용되는 모수벡터를 나타낸다. 일반적으로 가상관행렬은 적률법 또는 $\boldsymbol{\beta}$의 현재값을 사용하는 반복추정법에 의해 추정되는 미지의 행렬이다. 가상관행렬은 반복측정되는 자료의 특징을 고려하여 선택할 수 있다. 만약 가상관행렬이 $R_i(\boldsymbol{\rho}) = I_{t_i}$이면 GEE는 독립추정방정식 (independence estimating equation)이 된다. 모수벡터 $\boldsymbol{\rho}$의 형태에 따라 다양한 가상관행렬이 만들어지는데 교환가능행렬 (exchangeable matrix), 정상 1차 자기상관행렬 그리고 비모수적 행렬 등이 일반적으로 사용되고 있다. 자세한 내용은 Yeo 등 (2008), Liang과 Zeger (1986) 등에 설명되어 있다. 각 가상관행렬에 대한 모수 $\boldsymbol{\rho}$와 ϕ의 추정값은 Liang과 Zeger (1986)에 설명되어 있다. 여기서는 시계열 자료를 설명하는데 많이 활용되고 있으며 본 논문에서 사용할 정상 1차 자기상관행렬만을 소개한다.

$$R_i(\rho) = \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{t_i-1} \\ \rho & 1 & \rho & \cdots & \rho^{t_i-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{t_i-3} \\ & & & \vdots & \\ \rho^{t_i-1} & \rho^{t_i-2} & \rho^{t_i-3} & \cdots & 1 \end{pmatrix} \tag{2.4}$$

선택된 가상행렬 $R_i(\boldsymbol{\rho})$에 대해 GEE 방법에 의한 추정값 $\hat{\boldsymbol{\beta}}$와 $\hat{\boldsymbol{\rho}}$는 다음 추정방정식의 해이다.

$$\boldsymbol{S}(\boldsymbol{\beta}, \boldsymbol{\rho}) = \sum_{i=1}^{m} \frac{\partial \boldsymbol{\mu}_i^t}{\partial \boldsymbol{\beta}} V_i(\boldsymbol{\beta}, \boldsymbol{\rho})^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}. \tag{2.5}$$

GEE를 이용하여 구한 모수들의 추정량은 연결함수가 정확할 때 일치추정량이 되고 근사적으로 정규분포를 따르게 된다.

3. 일반화커널추정방정식 및 소지역 실업률 추정

소지역 추정 관련 모수적 모형이 완성단계 또는 한계에 이르렀다는 견해가 많아 모수적 모형을 넘어선 다른 방법을 통해 모형의 정확도를 향상시키려는 시도가 이루어지고 있다. 본 절에서는 경제활동인구조사처럼 여러 개의 소지역들을 같은 조사기간 동안 같은 조사시점에서 관측하여 얻은 패널자료 형태의 관측값을 갖는 반응변수를 모형화하기 위해 통계적 학습이론에서 많이 활용되고 있는 커널기법을 GEE에 적용하여 GEE의 비모수 버전인 GKEE를 유도하고 소지역 실업률의 추정에 활용하는 방법을 설명한다. 한편 소지역 추정을 위해 사용되는 모형은 크게 지역수준모형 (area level model)과 단위수준모형 (unit level model)으로 나누어진다. 소지역 실업률의 추정을 포함한 소지역 추정의 많은 경우 소지역의 각 단위들에 대한 보조정보를 알기 어렵기 때문에 본 논문에서처럼 지역수준모형을 사용한다.

3.1. 일반화커널추정방정식

이제 일반적인 소지역 추정을 위해 사용 가능한 GKEE 방법을 좀더 구체적으로 살펴보자. 본 논문에서는 m개의 소지역들을 같은 조사시점에서 같은 조사기간 동안 관측하여 얻은 패널자료를 이용하기 때문에 2절과는 달리 $t_1 = \cdots = t_m = T$의 경우를 생각한다. 2절에서처럼 y_{ij}는 i번째 소지역을 시간에 따라 반복측정할 때 j번째 시점에서 관측된 반응변수이다. 이때 $i = 1, \cdots, m$, $j = 1, \cdots, T$. 그리고 $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{iT})^t$는 i번째 소지역에 대한 반응변수들의 벡터를, $\boldsymbol{\mu}_i = (\mu_{i1}, \cdots, \mu_{iT})^t$는 대응되는 평균벡터를, $\boldsymbol{x}_{ij} = (x_{ij1}, \cdots, x_{ijp})^t$는 i번째 소지역에 대해 j번째 시점에서 이용 가능한 보조변수 또는 설명변수들의 벡터를 나타낸다. 소지역내의 각 관측값들 사이에 상관관계가 존재할 뿐만 아니

라 소지역간의 관측값들 사이에도 상관관계가 존재하기 때문에 통계적 학습이론의 커널기법을 적용하면 각 주변 반응변수의 기대값 $E(y_{ij}) = \mu_{ij}$와 설명변수벡터 $\boldsymbol{x}_{ij}$간의 관계식은 커널함수 $K(\cdot,\cdot)$를 사용하여 $\eta_{ij} = g(\mu_{ij}) = \sum_{k=1}^{m}\sum_{l=1}^{T}\alpha_{kl}K(\boldsymbol{x}_{ij},\boldsymbol{x}_{kl}) + b$로 표현할 수 있다. 커널함수 $K(\cdot,\cdot)$는 설명변수들의 선형항과 비선형항을 동시에 반영하기 위해 선형 커널함수와 가우시안 커널함수를 합한 것으로서 다음과 같이 정의된다.

$$K(\boldsymbol{x}_{ij},\boldsymbol{x}_{kl}) = \boldsymbol{x}_{ij}^t\boldsymbol{x}_{kl} + e^{-\frac{\|\boldsymbol{x}_{ij}-\boldsymbol{x}_{kl}\|^2}{2\sigma^2}}, \tag{3.1}$$

여기서 σ는 커널모수이다. 커널기법에 대한 개념 및 응용은 참고문헌 Hwang (2010, 2011), Hwang과 Shim (2012), Shim과 Hwang (2012, 2013), Vapnik (1995) 등에 설명되어 있다.

GKEE의 유도를 위해 관심 변수, 모수 및 커널행렬들을 벡터와 행렬을 사용하여 표현하고자 한다. 먼저 관심 모수와 변수들은 $\boldsymbol{\alpha}_i = (\alpha_{i1},\cdots,\alpha_{iT})^t$, $\boldsymbol{\alpha} = (b,\boldsymbol{\alpha}_1^t,\cdots,\boldsymbol{\alpha}_m^t)^t$, $\boldsymbol{\mu}_i = (\mu_{i1},\cdots,\mu_{iT})^t$, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^t,\cdots,\boldsymbol{\mu}_m^t)^t$, $\boldsymbol{\eta}_i = (\eta_{i1},\cdots,\eta_{iT})^t$, $\boldsymbol{\eta} = (\boldsymbol{\eta}_1^t,\cdots,\boldsymbol{\eta}_m^t)^t$, $\boldsymbol{y}_i = (y_{i1},\cdots,y_{iT})^t$, $\boldsymbol{y} = (\boldsymbol{y}_1^t,\cdots,\boldsymbol{y}_m^t)^t$로 표현된다. 커널행렬을 표현하기 위해 먼저 설명변수벡터 관련 두개의 집합 X와 X_i를 정의한다. X는 모든 설명변수벡터들의 집합을, X_i는 i번째 소지역 관련 설명변수벡터들의 집합을 나타낸다. 즉, $X = \{\boldsymbol{x}_{11},\cdots,\boldsymbol{x}_{mT}\}$, $X_i = \{\boldsymbol{x}_{i1},\cdots,\boldsymbol{x}_{iT}\}$이다. 그리고 X에 속하는 설명변수벡터들에 대한 커널함수값들이 원소인 $mT\times mT$ 행렬을 $K(X,X)$로 나타내고, X_i에 속하는 설명변수벡터들과 X에 속하는 설명변수벡터들에 대한 커널함수값들이 원소인 $T\times mT$ 행렬을 $K(X_i,X)$로 나타낸다. 그리고 커널행렬 K_0, K_1, K_{1i}를 다음과 같이 정의한다.

$$K_0 = \begin{pmatrix} 0 & \mathbf{0}_{mT}^t \\ \mathbf{0}_{mT} & K(X,X) \end{pmatrix},\ K_1 = (\mathbf{1}_{mT}, K(X,X)),\ K_{1i} = (\mathbf{1}_T, K(X_i,X)), \tag{3.2}$$

여기서 $\mathbf{0}_k$는 $k\times 1$ 영벡터를, $\mathbf{1}_k$는 모든 원소들이 1인 $k\times 1$ 벡터를 나타낸다.

따라서 $\boldsymbol{\eta}_i$와 $\boldsymbol{\eta}$는 각각 $\boldsymbol{\eta}_i = K_{1i}\boldsymbol{\alpha}$와 $\boldsymbol{\eta} = K_1\boldsymbol{\alpha}$로 표현되고 모수벡터 $\boldsymbol{\alpha}$의 추정값을 구하는 GKEE 방법의 준점수방정식은 다음과 같이 된다.

$$\begin{aligned} S(\boldsymbol{\alpha},\boldsymbol{\rho}) &= \sum_{i=1}^{m}\frac{\partial\boldsymbol{\mu}_i^t}{\partial\boldsymbol{\alpha}}V_i(\boldsymbol{\alpha},\boldsymbol{\rho})^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu}_i) - \lambda K_0\boldsymbol{\alpha} \\ &= \frac{\partial\boldsymbol{\mu}^t}{\partial\boldsymbol{\alpha}}V(\boldsymbol{\alpha},\boldsymbol{\rho})^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) - \lambda K_0\boldsymbol{\alpha} = \mathbf{0}, \end{aligned} \tag{3.3}$$

여기서 λ는 벌칙상수 (penalty parameter)이고 $V(\boldsymbol{\alpha},\boldsymbol{\rho})$는 $Cov(\boldsymbol{y}) = \phi V_i(\boldsymbol{\alpha},\boldsymbol{\rho})$를 만족하는 $V_i(\boldsymbol{\alpha},\boldsymbol{\rho})$로 이루어진 블록대각행렬이다. 다음 관계식

$$\frac{\partial\boldsymbol{\mu}^t}{\partial\boldsymbol{\alpha}} = \frac{\partial g(\boldsymbol{\mu})^t}{\partial\boldsymbol{\alpha}}\Big/\frac{\partial g(\boldsymbol{\mu})^t}{\partial\boldsymbol{\mu}} = K_1^tD \tag{3.4}$$

과 가종속변수 (working response variable) $\boldsymbol{y}^* = \boldsymbol{\eta} + D^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$를 이용하여 식 (3.3)을 $\boldsymbol{\alpha}$에 대하여 정리하면 다음과 같이 모수벡터 $\boldsymbol{\alpha}$를 구할 수 있다.

$$\boldsymbol{\alpha} = (K_1^tDV(\boldsymbol{\alpha},\boldsymbol{\rho})^{-1}DK_1 + \lambda K_0)^{-1}K_1^tDV(\boldsymbol{\alpha},\boldsymbol{\rho})^{-1}D\boldsymbol{y}^*. \tag{3.5}$$

이때 D는 $D = diag\{1/g'(\mu_{ij})\}$이다. 만약 g가 항등함수이면 D는 $mT\times mT$ 단위행렬 $D = I_{mT}$가 되고, g가 로짓함수이면 D는 $D = diag\{\mu_{ij}(1-\mu_{ij})\}$가 된다.

한편 $\boldsymbol{\alpha}$의 추정에 필요한 D와 $V(\boldsymbol{\alpha},\boldsymbol{\rho})$가 $\boldsymbol{\alpha}$를 포함하므로 식 (3.5)를 이용하여 한 번 만에 $\boldsymbol{\alpha}$를 구할 수는 없다. 따라서 이전 반복단계에서 구한 $\boldsymbol{\alpha}$의 추정값을 이용하여 D와 $V(\boldsymbol{\alpha},\boldsymbol{\rho})$를 구한 후 다시 $\boldsymbol{\alpha}$의 추정값을 구하는 반복알고리즘을 사용하여 $\boldsymbol{\alpha}$를 추정해야한다. 그 반복알고리즘은 다음과 같다.

1. $\boldsymbol{\alpha}$의 초기값을 $\boldsymbol{\alpha}^{(0)} = \mathbf{0}$로 놓는다.
2. $\boldsymbol{\alpha}^{(l)}$을 이용하여 D와 $V(\boldsymbol{\alpha}, \rho)$를 계산한다. 특히, 정상 1차 자기상관행렬을 사용하는 경우 $V(\boldsymbol{\alpha}, \rho)$를 다음과 같이 계산한다.

$$V(\boldsymbol{\alpha}, \rho) = A_i^{\frac{1}{2}} R_i(\rho) A_i^{\frac{1}{2}},\ R_i(\rho) = Corr(\boldsymbol{y}_i),\ A_i = diag\{Var(y_{ij})\},$$
$$\rho = \frac{1}{\phi(mT - m - df)} \sum_{i=1}^{m} \sum_{j \le T-1} r_{ij} r_{i,j+1},\ \phi = \frac{1}{mT - df} \sum_{i=1}^{m} \sum_{j=1}^{n} r_{ij}^2,$$
$$r_{ij} = \frac{y_{ij} - \mu_{ij}}{\sqrt{Var(y_{ij})}},\ df = trace(S),$$

   여기서 $S = K_1(K_1^t D V(\boldsymbol{\alpha}, \boldsymbol{\rho})^{-1} D K_1 + \lambda K_0)^{-1} K_1^t D V(\boldsymbol{\alpha}, \boldsymbol{\rho})^{-1} D$는 $\boldsymbol{\eta} = S\boldsymbol{y}^*$를 만족하는 모자행렬 (hat matrix)이다.
3. D와 $V(\boldsymbol{\alpha}, \rho)$를 이용하여 식 (3.5)의 $\boldsymbol{\alpha}^{(l+1)}$을 구한다.
4. $\|\boldsymbol{\alpha}^{(l+1)} - \boldsymbol{\alpha}^{(l)}\| < \epsilon$이 성립할 때까지 단계 2와 3을 반복한다.

이제 벌칙상수 λ와 커널모수 σ를 결정하는 모형선택 문제를 생각한다. GKEE 방법의 성능은 벌칙상수와 커널모수의 값에 영향을 받으므로 최적의 값을 선택하여야 한다. 즉 $\boldsymbol{\alpha}$를 추정하는 반복알고리즘의 단계 3에서 모형선택을 통하여 최적의 벌칙상수와 커널모수의 값을 결정해야 한다. 1차 Taylor 급수와 leaving-one-out lemma (Craven과 Wahba, 1979)를 이용하면 다음과 같이 일반화교차타당성 (generalized cross validation; GCV) 함수를 구할 수 있다.

$$GCV(\boldsymbol{\lambda}) = \frac{mT(\boldsymbol{y} - \boldsymbol{\mu})^t V(\boldsymbol{\alpha}, \boldsymbol{\rho})^{-1} (\boldsymbol{y} - \boldsymbol{\mu})}{(mT - df)^2}. \tag{3.6}$$

GKEE를 이용하여 $\hat{\boldsymbol{\alpha}}$를 구한 후 평균벡터 $\boldsymbol{\mu}_i$의 추정값 $\hat{\boldsymbol{\mu}}_i$를 다음과 같이 구할 수 있다.

$$\hat{\boldsymbol{\mu}}_i = g^{-1}(\hat{\boldsymbol{\eta}}_i) = g^{-1}(K_{1i}\hat{\boldsymbol{\alpha}}) \tag{3.7}$$

만약 연결함수 $g(\cdot)$가 항등함수이면 평균벡터의 추정값은 $\hat{\boldsymbol{\mu}}_i = K_{1i}\hat{\boldsymbol{\alpha}}$이고, 자연로그함수이면 추정값은 $\hat{\boldsymbol{\mu}}_i = \exp(K_{1i}\hat{\boldsymbol{\alpha}})$이며, 로짓함수이면 추정값은 $\hat{\boldsymbol{\mu}}_i = \frac{\exp(K_{1i}\hat{\boldsymbol{\alpha}})}{1+\exp(K_{1i}\hat{\boldsymbol{\alpha}})}$이 된다. 이때 벡터에 적용되는 연결함수의 역함수 $g^{-1}(\cdot)$와 지수함수 $\exp(\cdot)$는 편의상 각 성분별로 (componentwise) 적용되는 함수를 나타낸다고 가정한다.

3.2. 소지역 실업률 추정

GKEE를 이용하여 소지역 실업률을 추정하기 위해 다음과 같은 설계를 사용한다. 크기가 N인 모집단 U가 m개의 소지역 $U_i, i = 1, \cdots, m$로 구성되어 있다고 가정한다. 그러면 i번째 소지역 U_i의 크기를 N_i라고 할 때 $\sum_{i=1}^{m} N_i = N$가 성립한다. 그리고 크기 n인 표본을 추출하되 각 소지역에서 얻어진 표본크기를 n_i이라고 하면 $\sum_{i=1}^{m} n_i = n$가 성립한다. 한편 s_i를 i번째 소지역에서 표본으로 추출된 단위들의 집합, r_i를 이 소지역에서 표본으로 추출되지 않은 단위들의 집합이라 하면 $U_i = s_i \cup r_i, i = 1, \cdots, m$가 성립한다. 만약 y_{ij}가 i번째 소지역을 시간에 따라 반복측정할 때 j번째 시점에서의 실업자수를 나타내는 반응변수이면 소지역 실업률의 추정을 위해 실제로 사용되는 반응변수는 $y_{ij}^* = y_{ij}/n_i$이다. 왜냐하면 반응변수에 해당하는 경제활동인구 중 실업자의 수는 이항분포를 따른다고 가정하고 이에 따른 연결함수로는 이진자료 분석에서 일반적으로 많이 사용하고 있는 로짓함수를 적용하기 때문이다. 이때 n_i는 조사시점에 따라 다를 수 있지만 본 연구에서는 편의상 같다고 가정한다.

따라서 패널자료 $\{(\boldsymbol{x}_{ij}, y_{ij}^*)\}_{i=1,j=1}^{m,T}$에 GKEE를 적용하여 조사시점을 고려한 소지역 실업률의 추정값을 구하면 다음과 같이 된다.

$$\hat{\mu}_{ij}^{UE} = \frac{y_{ij} + (N_i - n_i) \times \hat{\mu}_{ij}}{N_i},\ i = 1, \cdots, m,\ j = 1, \cdots, T, \tag{3.8}$$

여기서 $\hat{\mu}_{ij}$은 $\hat{\boldsymbol{\mu}}_i = \frac{\exp(K_{1i}\hat{\boldsymbol{\alpha}})}{1+\exp(K_{1i}\hat{\boldsymbol{\alpha}})}$의 j번째 원소이다.

4. 실험연구

이 절에서는 조사시점을 고려한 소지역 실업률 추정에 사용 가능한 GKEE 방법의 성능을 살펴보기 위해 먼저 모의실험을 수행하여 GEE 방법, GLM과 성능을 비교한다. 그리고 2005년 1월부터 12월까지 경상남도 및 울산광역시의 25개 시군구의 경제활동인구조사 자료에 GKEE 방법, GEE 방법 및 GLM을 적용하여 시군구의 소지역 실업률을 추정한다. 먼저, 세 가지 방법의 성능을 비교할 수 있는 실제 자료를 구하는 것이 너무 어렵기 때문에 모의실험을 통하여 방법들을 비교하기로 한다. 모의실험에서는 2005년 경상남도 및 울산광역시의 25개 시군구의 경제활동인구조사와 비슷한 구조를 갖도록 모의실험을 위한 모집단을 설정하고 모의실험을 수행한다. 본 논문에서는 소지역의 각 단위들에 대한 보조정보를 알기 어렵기 때문에 지역수준모형을 사용하고 정상 1차 자기상관행렬을 가상관행렬로 사용한다.

4.1. 모의실험

모의실험의 과정을 세 단계로 나누어 설명하고자 한다. 첫 번째 단계는 각 소지역의 각 시점에서의 설명변수들의 값들을 생성하는 단계이고, 두 번째 단계는 각 소지역의 각 시점에서의 반응변수의 값을 생성하는 단계이며, 세 번째 단계는 생성된 모의실험 자료에 GKEE 방법, GEE 방법 및 GLM을 적용하여 얻은 추정결과를 비교하는 단계이다. 본 모의실험에서 소지역의 개수와 조사시점의 개수를 각각 20과 12로 설정하고, 관련 설명변수의 개수를 2로 설정한다. 그리고 각 소지역에 매 시점 1,000개의 단위가 있으며 각 소지역으로부터 매 시점 표본으로 100개의 단위를 추출한다고 가정한다.

먼저 첫 번째 단계인 각 소지역의 각 시점에서의 설명변수들의 값을 생성하는 과정을 설명하면 다음과 같다. 각 소지역에 매 시점 1,000개의 단위가 있기 때문에 매 시점 각 소지역의 단위들에 대한 보조정보를 나타내는 설명변수들의 값으로 균등분포 $U(0,1)$를 따르는 1,000개의 임의수 (random number) $x_{ij1}^k, x_{ij2}^k, i = 1, \cdots, 20,\ j = 1, \cdots, 12,\ k = 1, \cdots, 1,000$,를 생성하여 이들의 평균값들을 각 소지역의 설명변수들의 값으로 사용한다. 즉, 각 단위에 대한 설명변수의 값은 사실 알 수 없고 각 소지역에 대한 전체적인 특성만 알고 있다고 가정하여 지역수준모형을 사용하기 때문에 각 소지역의 설명변수들의 값은 $x_{ij1} = \frac{1}{1,000}\sum_{k=1}^{1,000} x_{ij1}^k$, $x_{ij2} = \frac{1}{1,000}\sum_{k=1}^{1,000} x_{ij2}^k$가 된다. 소지역 실업률 추정의 경우 x_{ij1}, x_{ij2}는 센서스 자료, 지역의 행정자료 및 인근 소지역에 관한 통계정보 등을 통해서 알 수 있는 각 소지역의 특성을 나타내는 보조정보에 해당된다.

이제 두 번째 단계인 각 소지역의 각 시점에서의 반응변수의 값을 생성하는 과정을 설명하면 다음과 같다. 먼저 다음과 같은 혼합효과모형을 이용하여 관련 모수들인 η_{ij}, μ_{ij}들을 생성한다.

$$\begin{aligned}
\eta_{i1} &= -1.5 + b_i + \exp(-x_{i11} - x_{i12}),\ b_i \sim N(0,1),\ i = 1, \cdots, 20, \\
\eta_{ij} &= -1.5 + b_i + \exp(-x_{ij1} - x_{ij2}) + 0.4\eta_{i,j-1},\ i = 1, \cdots, 20,\ j = 2, \cdots, 12, \\
\mu_{ij} &= \frac{\exp(\eta_{ij})}{1 + \exp(\eta_{ij})},\ i = 1, \cdots, 20,\ j = 1, \cdots, 12.
\end{aligned} \tag{4.1}$$

매 시점 i번째 소지역에서 표본으로 추출된 100개의 단위들의 집합을 s_i로, 표본으로 추출되지 않은 900개의 단위들의 집합을 r_i로 나타낸다. 그리고 s_i에 속하는 단위들에 대한 설명변수들의 평균을

$x^s_{ij1} = \frac{1}{100}\sum_{k\in s_i} x^k_{ij1}$, $x^s_{ij2} = \frac{1}{100}\sum_{k\in s_i} x^k_{ij2}$로 나타낸다. 그러면 x^s_{ij1}와 x^s_{ij2}를 식 (4.1)에 대입하여 μ_{ij}를 구한 후 표본으로 추출된 100개의 단위들에 대응되는 반응변수의 값 $y_{ij} = 100 \times \mu_{ij}$를 얻는다. 소지역 실업률 추정의 경우 y_{ij}는 경제활동인구조사를 통해서 얻게되는 각 소지역의 시점별 실업자수에 해당된다.

마지막으로 세 번째 단계인 추정결과를 비교하는 단계를 설명하면 다음과 같다. 첫 번째 단계와 두 번째 단계를 통하여 생성된 모의실험 자료 $\{(x_{ij1}, x_{ij2}, y_{ij})\}_{i=1,j=1}^{20,12}$에 GKEE 방법, GEE 방법 및 GLM을 적용하여 i번째 소지역에서 j번째 시점에 r_i에 속하는 900개 단위들의 실업률의 추정값에 해당하는 $\hat{\mu}_{ij}$을 얻을 수 있다. 따라서 구하고자 하는 소지역 모비율의 추정값은 다음과 같이 된다.

$$\hat{\mu}^{UE}_{ij} = \frac{y_{ij} + 900 \times \hat{\mu}_{ij}}{1,000},\ i = 1, \cdots, 20,\ j = 1, \cdots, 12. \tag{4.2}$$

그리고 각 추정방법에 대해 소지역 모비율의 추정값의 성능을 비교하기 위해

$$RMSE = \sqrt{\frac{1}{20 \times 12}\sum_{i=1}^{20}\sum_{j=1}^{12}(\hat{\mu}^{UE}_{ij} - \mu_{ij})^2} \tag{4.3}$$

를 계산한다.

전 과정을 100번 반복하여 100개의 RMSE의 평균값을 구한 것이 Table 4.1에 주어진다. 이때 괄호 안의 것은 표준오차이다. 제안된 GKEE 방법이 다른 두 방법보다 우수한 성능을 보이는 것을 알 수 있다.

Table 4.1 RMSE results for simulation study

Method	GKEE	GEE	GLM
RMSE	**9.9369**	12.9888	11.6792
	(0.0044)	(0.0012)	(0.0311)

4.2. 경제활동인구조사 자료를 이용한 실업률 추정

통계청에서는 매월 전국 약 32,000 표본가구 내에 거주하는 만 15세 이상인 사람들을 대상으로 경제활동 및 취업 여부 등과 같은 경제활동 관련 현황을 파악하기 위해 경제활동인구조사를 실시하고 있다. 경제활동인구조사는 1개 특별시, 6개 광역시 또는 9개 도 단위에 대한 통계 산출을 목적으로 표본설계가 되어 있다. 그러나 지방자치제도가 정착되면서 이제는 특별시, 광역시 또는 도 단위 뿐만 아니라 시군구 등과 같은 소지역 단위 통계에 대한 수요가 증폭되고 있다. 따라서 기존의 경제활동인구조사 자료를 기반으로 어느 정도 정교한 시군구별 소지역 통계 작성이 가능한 것인지에 대해서는 국가적인 차원에서 관심이 높은 연구영역이며, 이와 관련된 다양한 연구가 활성화 되는 것이 필요하다. 이런 관점에서 Yeo 등 (2008)은 GLM과 GEE를 적용한 소지역 추정법을 제시하고 있다. 경제활동인구조사 자료를 기반으로 한 본 연구의 실증분석에서는 Yeo 등 (2008)에서 사용한 자료와 동일한 2005년 1월부터 12월까지 경상남도 및 울산광역시에 대한 경제활동인구조사 자료에 GLM, GEE, GKEE를 적용하여 소지역 추정방법에 따른 차이를 보여주고자 한다.

분석대상이 되는 경상남도는 10개의 시와 10개의 군으로 이루어져 있으며, 울산광역시는 4개의 구와 한 개의 군으로 구성되어 있다. 해당 지역명과 주변지역 등과 관련된 내용은 Yeo 등 (2008)을 참고하기 바란다. 경제활동인구조사에서는 조사대상인 15세 이상 성인들을 대상으로 성별, 학력수준, 연령, 경제활동여부, 실업자 여부에 대한 조사가 이루어진다. 본 실증분석에서는 각 소지역의 월별 실업률을 추정하는 문제를 다루고 있으며, 여기서 실업률은 전체 경제활동인구 중에서 실업자가 차지하는 비율을 의미한다. 지역수준모형을 적용하기 위한 지역특성 설명변수로는 Yeo 등 (2008)에서 처럼 해당 지역의 경

제활동인구에서의 남성비율, 평균학력수준, 평균연령, 행정구역형태와 함께 주변 지역의 평균실업률을 사용하였다. 여기서 평균학력수준은 중졸미만, 중졸, 고졸, 초대졸, 대졸, 대학원이상으로 구분하였으며 행정구역 형태는 광역시 내에서 시와 군 지역을 구분하기 위해 가변수를 사용하였다.

반응변수에 해당하는 경제활동인구 중 실업자의 수는 이항분포를 따른다고 가정하고 이에 따른 연결함수로는 이진자료 분석에서 일반적으로 많이 사용하고 있는 로짓함수를 적용하였다. Table 4.2는 2005년 1월부터 12월까지 시군구별 실업률을 추정방법별로 추정한 결과이다. 추정방법을 나타내는 Sample은 경제활동인구조사 자료에서 구한 각 소지역의 실업자수를 각 소지역의 조사대상자수로 나눈 단순 실업률을 의미한다. 여기서는 소지역별 실업률 모수를 알 수 없기 때문에 추정방법에 따른 효율성을 비교하는 작업을 수행하지 못하였고, 추정방법에 따른 추정상의 차이점만을 보여주고 있다는 점을 참고하기 바란다.

Table 4.2 Estimated unemployment rates

Area	Method	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
A	Sample	4.22	1.79	2.41	3.57	2.40	2.47	1.26	2.56	2.52	1.85	1.18	2.14
	GKEE	3.44	2.92	2.71	2.74	2.44	2.42	2.69	2.59	2.54	2.62	2.46	2.88
	GEE	2.52	2.87	2.92	3.22	3.18	2.93	3.12	2.78	3.07	2.63	2.76	2.61
	GLM	4.14	3.36	2.88	2.44	2.92	2.63	3.24	2.51	2.79	2.13	2.25	1.98
B	Sample	1.92	1.79	0.00	1.56	1.56	3.03	0.00	0.00	0.00	2.94	1.47	2.70
	GKEE	0.85	1.05	0.74	0.76	1.15	0.72	0.74	0.88	0.35	0.71	0.74	0.84
	GEE	1.30	1.36	1.35	1.36	1.39	1.30	1.24	1.16	1.28	1.22	1.22	1.24
	GLM	0.95	1.12	1.10	1.34	1.39	1.41	1	0.36	1.26	0.71	0.39	1.60
C	Sample	2.70	2.82	1.35	2.67	2.60	4.05	2.70	2.63	1.30	1.30	2.63	0.00
	GKEE	2.13	2.76	1.79	2.92	2.71	3.06	3.08	1.63	1.97	1.34	2.17	1.17
	GEE	1.94	1.97	2.01	2.01	2.00	1.88	1.94	1.94	2.01	1.95	1.94	1.95
	GLM	3.67	2.38	2.59	2.22	2.16	2.66	3.01	2.83	3.12	2.10	1.89	0.99
D	Sample	5.43	3.66	2.38	2.73	2.38	3.24	2.73	1.95	1.93	2.68	3.85	3.50
	GKEE	3.81	4.11	2.38	2.62	2.39	2.42	3.56	2.71	2.61	2.78	2.90	2.62
	GEE	2.83	2.81	2.85	2.75	2.71	3.03	2.96	2.86	2.67	2.90	2.89	2.79
	GLM	2.85	3.00	2.43	2.71	2.04	3.12	3.19	3.31	2.44	3.14	3.28	2.75
E	Sample	5.16	4.92	5.11	4.70	4.29	3.94	4.78	4.47	3.69	3.82	2.82	3.87
	GKEE	3.40	3.64	3.43	3.42	3.55	3.89	3.82	3.83	3.66	3.09	2.51	3.41
	GEE	3.57	3.73	3.73	3.65	3.71	3.67	3.67	3.70	3.71	3.55	3.45	3.50
	GLM	4.52	4.23	3.99	3.67	3.96	3.48	3.38	3.62	3.21	3.13	2.28	3.39
F	Sample	0.00	1.75	3.45	1.79	0.00	1.75	0.00	0.00	0.00	0.00	0.00	0.00
	GKEE	1.00	1.19	1.54	0.96	0.91	1.49	1.32	0.52	0.47	0.70	0.46	1.60
	GEE	1.06	1.17	1.17	1.09	1.11	1.13	1.12	1.13	1.13	1.17	1.15	1.14
	GLM	0.90	1.45	2.26	1.96	1.05	1.06	1.36	0.42	1.88	0.64	0.31	0.58
G	Sample	1.74	3.57	3.90	3.18	2.71	3.07	2.08	1.80	2.52	1.09	1.15	1.50
	GKEE	2.67	2.73	2.99	2.60	2.50	2.63	2.37	2.42	2.30	2.70	2.43	2.58
	GEE	3.29	3.43	3.50	3.47	3.46	3.19	3.31	3.32	3.19	3.21	3.14	3.19
	GLM	3.40	4.18	3.88	3.59	3.67	3.04	3.16	2.48	2.66	2.55	1.75	2.68
H	Sample	4.30	3.83	4.03	3.03	3.34	2.77	2.41	2.84	2.40	2.04	3.36	3.92
	GKEE	2.83	3.22	3.03	3.12	3.02	3.16	2.80	3.00	2.35	2.37	2.37	2.83
	GEE	2.99	3.17	2.91	2.93	2.90	3.09	3.09	2.99	2.93	2.90	3.02	2.84
	GLM	4.48	3.14	2.98	2.63	2.31	2.40	2.91	2.97	2.58	2.75	3.50	3.22
I	Sample	3.42	3.94	3.13	1.54	0.00	4.26	5.84	5.07	2.90	2.92	3.01	3.61
	GKEE	2.38	2.47	2.42	2.79	2.27	3.67	4.24	4.17	4.11	3.98	2.91	2.36
	GEE	2.47	2.46	2.50	2.41	2.36	2.45	2.54	2.51	2.60	2.64	2.66	2.54
	GLM	3.03	2.30	1.81	2.13	1.95	1.89	2.05	2.18	2.00	2.47	2.48	3.47
J	Sample	3.60	3.52	3.02	3.80	3.43	1.73	1.30	2.16	1.72	1.28	1.69	2.65
	GKEE	2.86	3.45	3.79	3.13	3.52	3.17	3.33	3.32	2.36	2.28	2.37	2.84
	GEE	3.51	3.75	3.85	3.69	3.64	3.64	3.63	3.56	3.58	3.44	3.25	3.26
	GLM	4.61	4.33	4.00	3.65	4.07	4.18	3.63	3.84	3.03	3.09	1.97	2.48
K	Sample	3.16	2.15	3.19	2.80	3.70	2.68	2.52	3.54	4.27	1.69	0.00	1.35
	GKEE	2.46	2.40	3.44	2.96	2.71	2.34	2.03	3.48	3.47	1.86	0.99	2.60
	GEE	2.01	2.01	1.98	2.06	2.01	2.07	2.12	2.08	2.07	2.06	2.09	1.97
	GLM	3.69	1.97	1.96	1.82	1.96	1.96	1.53	2.20	1.81	1.37	1.09	1.75
L	Sample	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	GKEE	0.62	0.67	0.86	0.77	0.27	0.26	0.26	0.33	0.60	0.52	0.65	0.69
	GEE	1.30	1.27	1.28	1.33	1.26	1.21	1.20	1.23	1.33	1.30	1.35	1.34
	GLM	1.24	1.13	0.94	1.17	1.40	1.67	0.95	0.68	1.47	0.88	0.60	1.10
M	Sample	4.82	2.44	3.11	1.80	0.63	0.63	1.92	3.90	1.94	3.33	4.79	3.12
	GKEE	3.40	3.29	2.98	2.27	1.72	1.81	2.04	3.08	3.18	2.60	2.54	3.19
	GEE	3.43	3.52	3.33	3.32	3.11	3.47	3.62	3.43	3.24	3.25	3.21	3.22
	GLM	4.91	3.74	2.92	2.85	2.39	3.02	3.74	4.86	2.78	3.65	4.23	3.50

Estimated unemployment rates (continued)

Area	Method	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
	Sample	1.98	2.73	2.96	2.62	2.24	2.90	4.85	3.59	4.22	3.18	2.26	2.27
N	GKEE	2.78	2.76	3.36	2.68	2.93	3.13	3.34	2.74	2.85	2.69	2.71	2.71
	GEE	2.33	2.27	2.25	2.13	2.10	2.15	2.22	2.27	2.14	2.13	2.12	2.17
	GLM	1.67	2.52	2.00	2.40	1.68	2.31	3.41	2.64	2.82	2.55	2.53	1.80
	Sample	2.86	2.94	4.88	5.00	2.44	0.00	0.00	2.56	2.70	0.00	2.50	2.94
O	GKEE	1.86	1.91	3.01	2.11	1.62	1.42	1.21	1.15	2.34	1.55	2.34	1.29
	GEE	1.40	1.33	1.35	1.40	1.55	1.57	1.58	1.56	1.55	1.43	1.46	1.49
	GLM	1.05	1.23	1.58	1.66	1.50	1.70	1.46	0.90	1.71	1.04	0.71	2.01
	Sample	6.25	4.68	3.66	2.61	4.46	3.50	3.46	4.12	3.53	3.51	1.91	3.21
P	GKEE	4.59	3.50	3.54	3.55	3.70	3.74	3.51	3.39	3.51	3.51	3.83	3.64
	GEE	3.23	3.34	3.32	3.27	3.17	3.06	3.20	3.21	3.18	3.08	2.99	3.20
	GLM	4.52	3.96	3.74	3.32	3.60	3.52	3.18	3.37	2.86	2.47	1.58	2.59
	Sample	1.53	1.19	0.78	1.16	0.39	0.00	1.85	1.52	1.89	2.23	2.25	2.46
Q	GKEE	2.38	2.16	1.15	1.56	1.32	1.59	2.88	2.24	2.14	2.33	2.60	2.65
	GEE	2.64	2.68	2.67	2.78	2.77	2.82	2.83	2.89	2.86	2.65	2.79	2.67
	GLM	2.42	2.96	2.83	2.91	2.02	1.57	2.75	1.73	2.80	1.98	2.29	2.45
	Sample	2.53	1.99	1.29	1.96	3.31	6.72	3.15	1.89	0.97	4.59	1.77	2.55
R	GKEE	3.20	2.44	2.03	2.91	2.05	5.56	3.44	2.26	1.64	3.08	2.18	2.10
	GEE	3.88	3.89	3.70	3.66	3.46	3.62	3.58	3.48	3.39	3.58	3.59	3.48
	GLM	4.36	3.83	3.32	3.11	3.74	4.26	3.05	3.16	2.46	3.57	3.44	3.99
	Sample	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.96	1.79	0.00	2.35
S	GKEE	0.60	0.52	0.35	0.43	0.47	0.41	0.49	0.56	1.04	1.03	1.08	1.05
	GEE	1.25	1.24	1.22	1.20	1.15	1.23	1.24	1.18	1.19	1.13	1.21	1.20
	GLM	1.38	1.11	0.93	1.28	1.33	2.94	1.26	1.25	1.32	0.72	0.49	0.98
	Sample	3.46	4.44	3.50	3.26	2.84	3.57	4.23	3.29	3.13	4.38	4.48	2.66
T	GKEE	3.24	3.33	3.25	3.09	3.00	3.75	3.39	3.19	3.54	3.48	3.54	2.92
	GEE	3.26	3.29	3.22	3.17	3.23	3.63	3.50	3.35	3.34	3.35	3.32	3.31
	GLM	4.11	3.76	3.69	3.18	2.43	2.98	3.63	3.30	3.02	3.73	4.37	3.23
	Sample	0.85	2.42	2.34	4.03	4.03	2.61	3.67	1.89	3.57	1.15	1.08	0.94
U	GKEE	1.54	2.13	2.75	3.30	3.45	2.60	3.20	3.38	2.30	2.13	1.77	1.86
	GEE	2.16	2.02	1.99	2.14	2.09	2.15	2.04	2.11	2.03	2.00	2.05	1.94
	GLM	1.64	2.04	1.83	2.00	1.70	2.54	1.71	1.57	1.89	1.30	1.04	1.78
	Sample	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V	GKEE	0.35	0.34	0.53	0.36	0.52	0.46	0.79	0.70	1.34	1.21	1.17	0.58
	GEE	1.41	1.43	1.52	1.55	1.56	1.51	1.44	1.42	1.47	1.38	1.36	1.37
	GLM	1.27	1.02	0.63	1.06	1.55	1.61	1.12	0.64	1.44	0.93	0.52	2.08
	Sample	1.67	1.64	0.00	0.00	1.54	3.08	3.08	1.79	1.72	0.00	1.64	0.81
W	GKEE	1.39	1.31	1.48	1.16	1.50	2.21	2.36	1.31	1.40	1.10	1.07	2.22
	GEE	1.68	1.69	1.72	1.72	1.52	1.54	1.57	1.60	1.63	1.60	1.65	1.70
	GLM	2.96	1.71	1.38	1.65	1.81	2.57	1.73	3.23	2.31	1.57	1.36	0.97
	Sample	3.45	3.23	3.33	4.00	3.70	0.00	0.00	0.00	3.45	0.00	0.00	1.75
X	GKEE	3.41	3.21	3.18	4.15	2.45	0.50	0.71	1.49	2.32	1.11	0.61	1.92
	GEE	2.81	2.30	2.17	2.18	2.38	2.57	2.27	2.35	2.29	2.42	2.36	2.48
	GLM	2.22	3.40	5.17	2.32	1.84	1.55	2.16	0.97	2.22	2.82	2.72	3.79
	Sample	3.33	0.00	0.00	1.27	1.27	1.32	0.00	0.00	1.30	1.28	1.37	1.63
Y	GKEE	1.04	0.75	0.91	0.90	1.12	1.24	1.35	0.84	0.93	1.52	1.36	1.13
	GEE	1.33	1.32	1.32	1.47	1.43	1.43	1.37	1.41	1.55	1.53	1.52	1.48
	GLM	1.62	1.23	1.38	1.42	1.60	1.79	1.11	0.90	1.63	1.24	0.820	1.45

5. 결론

기존의 소지역 추정에 대한 대부분의 연구들은 특정 시점에서의 추정에 국한 되어 왔다. 그러나 경제활동인구조사와 같은 많은 국가통계 작성을 위한 조사들에서는 월, 분기 또는 연 단위로 조사를 수행하기 때문에 결과적으로 패널자료를 생산하게 된다. 따라서 이를 고려한 소지역 추정방법의 개발이 필요하다. 본 논문에서는 소지역의 실업률 관련 패널자료의 분석을 위해 유용하게 사용되고 있는 GEE의 비모수 버전인 GKEE 방법을 제안하였다. 모의실험을 통하여 제안된 GKEE 방법이 GEE 방법과 GLM 보다 더 좋은 추정결과를 보여주는 것을 알 수 있었다. 그리고 2005년 1월부터 12월까지 경상남도 및 울산광역시의 25개 시군구의 경제활동인구조사 자료에 세 가지 방법을 적용하여 해당 소지역의 월별 실업률을 추정하여 보았다. 그러나 소지역 실업률 모수를 모르기 때문에 추정의 성능을 제대로 평가하지는 못하고 있다. 향후 보다 의미 있는 효율성 비교를 위해서는 소지역별 실업률을 파악할 수 있는 자료를 대상으로 추정결과를 비교하는 연구가 필요할 것이다.

References

Chung, Y. S., Lee, K. and Kim, B. C. (2003). Adjustment of unemployment estimates based on small area estimation in Korea. *Survey Methodology*, **29**, 45-52.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, **31**, 377-403.

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.

Hwang, C. (2010). Support vector quantile regression for longitudinal data. *Journal of Korean Data & Information Science Society*, **21**, 309-316.

Hwang, C. (2011). Asymmetric least squares regression estimation using weighted least squares support vector machine. *Journal of Korean Data & Information Science Society*, **22**, 995-1005.

Hwang, C. and Shim, J. (2012). Mixed effects least squares support vector machine for survival data analysis. *Journal of Korean Data & Information Science Society*, **23**, 739-748.

Jeong, S. and Shin, K. (2012). Small area estimation via nonparametric mixed effects model. *The Korean Journal of Applied Statistics*, **25**, 457-464.

Khoshgooyanfard, A. and Monazzah, M. T. (2006). A cost effective strategy for provincial unemployment estimation: A small area approach. *Survey Methodology*, **32**, 105-114.

Kim, Y. and Choi, H. (2004). Small area estimation techniques based on logistic model to estimate unemployment rate. *Communications of the Korean Statistical Society*, **11**, 583-595.

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

Marker, D. A. (1999). Organization of small area estimators using a generalized linear regression framework. *Journal of Official Statistics*, **15**, 1-24.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society A*, **135**, 370-384.

Noble, A., Haslett, S. and Arnold, G. (2002). Small area estimation via generalized linear models. *Journal of Official Statistics*, **18**, 45-60.

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

Pereira, L. N., Mendes, J. M. and Coelho, P. S. (2013). Model-based estimation of unemployment rates in small areas of Portugal. *Communications in Statistics - Theory and Methods*, **42**, 1325-1342.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, **28**, 40-68.

Rao, J. N. K. (2003). *Small area estimation*, John Wiley & Sons, New Jersey.

Salvati, N., Chandra, H., Ranalli, M. G. and Chambers, R. (2010). Small area estimation using a nonparametric model-based direct estimator. *Computational Statistics and Data Analysis*, **54**, 2159-2171.

Shim, J. and Hwang, C. (2012). M-quantile kernel regression for small area estimation. *Journal of Korean Data & Information Science Society*, **23**, 749-756.

Shim, J. and Hwang, C. (2013). Expected shortfall estimation using kernel machines. *Journal of Korean Data & Information Science Society*, **24**, 12-20.

Ugarte, M. D., Goicoa, T., Militino, A. F. and Sagaseta-López, M. (2009). Estimating unemployment in very small areas. *SORT*, **33**, 49-70.

Vapnik, V. N. (1995). *The nature of statistical learning theory*, Springer, New York.

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

Yeo, I., Son, K. and Kim, Y. (2008). Small area estimation via generalized estimating equations and the panel analysis of unemployment rates. *The Korean Journal of Applied Statistics*, **21**, 665-674.

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.

Journal of the Korean Data & Information Science Society
2013, **24**(6), 1199–1210

http://dx.doi.org/10.7465/jkdi.2013.24.6.1199
한국데이터정보과학회지

Generalized kernel estimating equation for panel estimation of small area unemployment rates[†]

Jooyong Shim[1] · Youngwon Kim[2] · Changha Hwang [3]

[1]Department of Data Science, Inje University
[2]Department of Statistics, Sookmyung Women's University
[3]Department of Applied Statistics, Dankook University

Received 29 June 2013, revised 4 August 2013, accepted 23 August 2013

Abstract

The high unemployment rate is one of the major problems in most countries nowadays. Hence, the demand for small area labor statistics has rapidly increased over the past few years. However, since sample surveys for producing official statistics are mainly designed for large areas, it is difficult to produce reliable statistics at the small area level due to small sample sizes. Most of existing studies about the small area estimation are related with the estimation of parameters based on cross-sectional data. By the way, 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 this type of panel data. In this paper, we derive the generalized kernel estimating equation which can model time-dependency among response variables and handle repeated measurement or panel data. We compare the proposed estimating equation with the generalized linear model and the generalized estimating equation through simulation, and apply it to estimating the unemployment rates of 25 areas in Gyeongsangnam-do and Ulsan for 2005.

Keywords: Generalized estimating equation, generalized kernel estimating equation, generalized linear model, kernel technique, panel data, panel estimation, small area estimation.

[†] This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2011-0021125).

[1] Adjunct professor, Department of Data Science, Inje University, Kyungnam 621-749, Korea.
[2] Professor, Department of Statistics, Sookmyung Women's University, Seoul 140-742, Korea.
[3] Professor, Department of Applied Statistics, Dankook University, Gyeonggido 448-160, Korea. E-mail: chwang@dankook.ac.kr