응용통계연구 (2012)
25(2), 293–300
DOI: http://dx.doi.org/10.5351/KJAS.2012.25.2.293

이단계 소지역추정

이상은[1] · 신기일[2]

[1]경기대학교 응용정보통계학과, [2]한국외국어대학교 통계학과

(2011년 12월 30일 접수, 2012년 1월 30일 수정, 2012년 3월 6일 채택)

요 약

지역 또는 도메인에 작은 크기의 표본이 배정되어 추정의 정도가 낮을 때 사용하는 통계적 기법인 소지역추정에 관한 많은 연구가 진행되고 있다. 소지역추정에 사용되는 자료는 단위수준자료(unit level data)와 지역수준자료(area level data)로 분류된다. 본 논문에서는 단위수준자료를 이용하여 소지역추정을 실시한 후 얻어진 추정값에 공간통계분석기법을 도입하여 최종적인 소지역추정값을 얻는 이단계 소지역추정법을 제안하였다. 제안된 소지역추정법은 단위수준자료가 갖고 있는 정보와 지역수준자료가 갖고 있는 공간정보를 모두 이용하는 방법으로 추정의 정도를 높일 수 있는 새로운 방법이다. 본 논문에서는 경제활동인구조사 자료를 이용한 모의실험을 통해 이단계 소지역추정법의 우수성을 확인하였다.

주요용어: 단위수준자료, 지역수준자료, 로짓회귀모형, 로짓혼합모형, 공간추정량.

1. 서론

전국이나 시도와 같이 광역지역 단위가 아니라 시군구나 동읍면과 같이 추정범위가 작은 경우의 추정 문제를 다루는 소지역추정에 대해 활발한 연구가 진행되고 있다 (Rao, 2003).

단위수준자료, 특히 실업자 수 자료는 실업률 자체가 낮기 때문에 정확한 소지역추정에 어려움이 있다. 최근 통계청에서는 이러한 소지역추정의 어려움을 극복하기 위하여 20만 가구 이상의 표본을 추출하여 조사하고 있다. 이 자료는 국내의 미진한 소지역추정에 중요한 기초 자료로 사용될 수 있을 것이다.

실업자 수와 같이 이항자료 추정을 위해 흔히 사용되는 소지역추정 방법은 로지스틱회귀모형(logistic regression model)과 로지스틱혼합모형(logistic mixed model)을 이용한 추정법이다. 로지스틱회귀형태의 소지역추정법에 관한 많은 연구가 진행되었으며 이에 관한 내용은 최영아 (2004), 여인권 등 (2008), Hwang과 Shin (2011)을 참조하기 바란다.

그러나 이러한 방법은 단위수준이 갖고 있는 정보만을 이용하여 소지역을 추정하는 방법으로 지역수준의 정보를 사용하고 있지 않다. 물론 로지스틱혼합모형의 경우에는 랜덤효과 부문에 지역이 포함되어 있으므로 지역수준정보를 사용하고 있다고 할 수 있으나 정확한 의미의 지역수준정보를 사용한다고 볼 수 없다. 널리 사용되고 있는 지역수준정보로는 시간적인 관계를 나타내는 자기상관관계와 공간적인 관계를 나타내는 공간상관관계가 있다. 자기상관관계를 분석하기 위해서는 이를 모형화할 수 있는 충분한

이 논문은 2009년도 정부(교육과학기술원)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임 (2009-0072926).
[2]교신저자: (449-791) 경기도 용인시 모현면 왕산리 산 89, 한국외국어대학교 통계학과, 교수.
E-mail: keyshin@hufs.ac.kr

과거 자료가 얻어져야 하므로 이 조건이 맞지 않는 경우 실제 자료 분석에 적용하기에는 다소 무리가 따른다. 반면 공간상관관계는 횡단면 자료에서 얻을 수 있는 정보이므로 분석에 적용하기가 용이하다. 다만 유의한 결과를 얻기 위해서는 유의한 공간상관관계가 존재하여야 한다. 신기일 등 (2007)은 경기지역 자료를, 이강석과 신기일 (2008)은 전라북도 자료를 이용하여 공간소지역추정량을 연구하였다.

본 논문에서는 경제활동상태 파악에서 중요하게 다루어지는 실업자 수의 소지역추정을 다루므로 이러한 자료에 널리 사용되는 로지스틱회귀추정량과 로지스틱혼합추정량을 고려하였다. 이단계 소지역추정법은 먼저 로지스틱회귀추정량 또는 로지스틱혼합추정량을 이용하여 소지역추정을 실시한다. 이때 얻어진 소지역추정값을 지역수준 자료로 활용한 후 이 결과에 공간상관관계를 접목하여 소지역추정량을 얻는 방법이다.

본 논문의 구성은 다음과 같다. 먼저 2절에서 단위수준자료가 이항자료인 경우의 소지역추정에 흔히 사용하는 로지스틱회귀추정량과 로지스틱혼합추정량을 설명하였다. 3절에서는 이단계 소지역추정량을 위한 공간추정량을 설명하였으며 이를 종합하여 얻어진 이단계 소지역추정법을 설명하였다. 4절에서는 경제활동자료를 이용한 모의실험이 수행되었으며, 모의실험을 통해 제안된 방법의 우수성을 확인하였다. 5절에 결론이 있다.

2. 일단계 소지역추정량

이단계 소지역추정법은 다음의 두 단계로 이루어진다. 먼저 일단계에서는 단위수준자료를 이용하여 각 소지역의 총계를 추정한다. 일단계 소지역추정에 사용되는 추정량은 직접추정량, 회귀추정량 등 다양한 방법이 사용될 수 있으나 본 논문에서는 특히 이항 변수인 실업자 수 자료에 우수한 결과를 주는 것으로 알려진 로지스틱회귀모형과 로지스틱혼합모형을 살펴보았다.

이단계 추정은 일단계에서 얻어진 시군구별 소지역추정값을 이용하여 공간추정법을 도입한 후 최종적인 소지역추정값을 얻는다. 이렇게 얻어진 두개의 추정량이 공간로지스틱회귀추정량과 공간로지스틱혼합추정량이 된다. 일반적인 실제 자료 분석에서는 각 소지역 i와 그 소지역에 포함된 j번째 자료의 설계가중치 w_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$가 사용되지만 본 논문에서는 자료 분석과 모의실험을 간단히 하기위해 설계가중치 $w_{ij} = 10$을 사용하였다. 물론 이 가중치 값은 분석 및 추정량의 성능 비교에 영향을 주지 않는다. 이 절에서는 일단계 소지역추정에서 사용되는 각 모형을 간단히 살펴보았다.

2.1. 로지스틱회귀추정량($\hat{Y}^{LOGIT}$)

일반적으로 반응변수(Y)가 오직 두 개의 범주만을 갖는 이항자료이며 p개의 설명변수 $X_1, \ldots, X_p$가 있을 때 이항반응의 성공확률 $P(Y = 1|x_1, \ldots, x_p) = p(x)$에 대한 로지스틱회귀모형(logistic regression model)은

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{2.1}$$

으로 정의된다. 이를 다른 형태로 나타내면

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}$$

와 같다. 여기서 설명변수의 일부는 질적 변수일 수도 있다. 본 연구에서 사용될 자료는 경제활동인구조사 자료이다. 이 자료에는 이항자료인 경제활동 여부 즉 취업 여부를 나타내는 종속변수와 설명변수에 해당되는 학력(X_1), 연령(X_2), 비농가여부(X_3), 성별(X_4) 등이 포함되어 있다. n개의 소지역이

있고 i번째 소지역에는 n_i개의 관측값이 존재한다. 이제 i번째 소지역의 j번째 실업 여부에 대한 관측값을 y_{ij}라 하면, p_{ij}추정값은 $\log(\hat{p}_{ij}/(1-\hat{p}_{ij})) = \hat{\beta}_0 + \hat{\beta}_1 x_{1ij} + \cdots + \hat{\beta}_p x_{pij}$로 구해진다. 여기서 $\hat{p}_{ij} = \hat{P}(y_{ij} = 1|x_{1ij}, \ldots, x_{pij})$, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$이고 $\sum_{j=1}^{n_i} w_{ij}\hat{p}_{ij}$를 구하면 이 값이 i지역 실업자 수 Y_i에 대한 추정값이 된다. 여기서 w_{ij}는 설계가중치이며 본 논문에서는 $w_{ij} = 10$을 사용하였다.

2.2. 로지스틱혼합추정량($\hat{Y}^{LMM}$)

2.1절의 로지스틱회귀모형에서 설명변수로 설명되어지는 부분인 체계적 성분(systematic component)에 랜덤효과(random effect)를 추가로 고려한 모형이 로지스틱혼합모형(logistic mixed model)이며 모형은 다음과 같이 정의된다.

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + v \tag{2.2}$$

또는

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + v)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + v)},$$

여기서 v는 소지역 차이에 따른 효과로 이를 랜덤효과로 간주한다. 식 (2.3)의 혼합모형을 선택하게 되면 랜덤효과에 대한 분산을 분리하여 계산할 수 있어, 소지역별 추정값들이 서로 크게 차이가 날 경우에는 분산을 줄이는 효과가 있다. 추정된 모형식은 다음과 같다.

$$\log\left(\frac{\hat{p}_{ij}}{1-\hat{p}_{ij}}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_{1ij} + \cdots + \hat{\beta}_p x_{pij} + \hat{v}_i,$$

여기서 $\hat{p}_{ij} = \hat{P}(y_{ij} = 1|x_{1ij}, \ldots, x_{pij})$, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$이고 $\sum_{j=1}^{n_i} w_{ij}\hat{p}_{ij}$를 구하여 i지역 실업자 수 $\hat{Y}_i = \sum_{j=1}^{n_i} w_{ij}\hat{p}_{ij}$로 소지역추정값을 얻는다. 여기서 w_{ij}는 설계가중치이며 본 논문에서는 $w_{ij} = 10$을 사용하였다.

3. 이단계 소지역추정량

먼저 각 소지역별로 단위수준자료인 이항자료가 얻어졌다고 가정하자. 즉 자료는 성공이면 "1", 실패이면 "0"이다. 본 연구에서 사용된 자료는 실업자 수이므로 실업자이면 "1" 그리고 실업자가 아닌 취업자 또는 비경제활동자이면 "0"이 된다.

3.1. 일단계 소지역추정량

이제 각 소지역별로 로지스틱회귀모형과 로지스틱혼합모형을 적용하여 얻어진 소지역추정량을 다음과 같이 표시하자. 이때 비교를 위한 직접추정량, $\hat{Y}^{DE}$도 구하였다.

$$\text{직접추정량} : \hat{Y}^{DE}$$
$$\text{로지스틱회귀추정량} : \hat{Y}^{LOGIT}$$
$$\text{로지스틱혼합추정량} : \hat{Y}^{LMM}$$

여기서 로지스틱회귀추정량은 식 (2.1)을 로지스틱혼합추정량은 식 (2.2)를 이용하여 얻어진다.

표 **3.1.** 전라북도 이웃정보시스템

소지역		이웃 소지역 번호	소지역		이웃 소지역 번호
행정구역코드	번호		행정구역코드	번호	
35010	1	3, 6, 7	35310	7	1, 3, 4, 6, 8, 10
35020	2	3, 6	35320	8	7, 9, 10
35030	3	1, 2, 6, 7	35340	9	5, 8, 10
35040	4	6, 7, 10, 11, 12, 13	35350	10	4, 5, 7, 8, 9, 11
35050	5	5, 9, 11	35360	11	4, 5, 10
35060	6	1, 2, 3, 4, 7, 13	35370	12	4
			35380	13	4, 6, 12

3.2. 이웃정보시스템을 이용한 공간추정량

3.2.1. 이웃정보시스템 공간통계자료는 크게 지리통계자료(geostatistical data)와 격자자료(lattice data)로 나누어지며 소지역추정에는 일반적으로 격자자료가 사용된다. 격자자료 분석의 가장 중요한 단계는 격자자료의 공간상관관계를 정의하기 위해 이웃정보시스템을 구축하는 것이다. 이웃을 어떤 방법으로 정의하는가에 따라 이웃정보시스템은 다양하게 나타나게 된다. 이웃을 정의하기 위한 기준으로, 경계공유 유무 기준, 거리 기준, 최근거리 기준 등이 있으며 정해진 방법의 각 이웃 값에 가중치를 차등하여 주어서 최종 이웃을 결정할 수도 있다. 이와 같이 여러 방법으로 이웃정보시스템을 결정할 수 있으며, 이웃정보시스템의 정의 및 비교에 관한 자세한 내용은 Cressie (1993)를 살펴보기 바란다. 또한 Rao (2003)에서는 SAR(spatial autoregression) 모형을 이용한 공간소지역추정을 간단히 설명하였다. 또한 국내 적용사례 및 각 지역의 공간통계 연구는 이상은 (2006), 김재두 등 (2005), 이강석과 신기일 (2008)을 살펴보기 바란다. 이 연구 결과를 종합하면 이웃을 공유하는 시군구를 이웃으로 정하는 방법이 쉬우면서도 매우 효과적인 것으로 판단되었다. 이에 따라 본 논문에서는 경계를 공유하는 시군구를 이웃으로 정하였다. 물론 공간상관관계는 각 자료마다 다르기 때문에 분석전에 공간상관관계를 파악할 수 있는 Moran's I를 구하여 공간상관관계를 확인하여야 한다. 즉 공간상관관계가 존재하여야만 공간추정량 사용이 의미가 있다.

3.2.2. 공간추정량 소지역 간의 공간상관관계를 이용한 SAR 모형은 다음과 같이 정의된다.

$$Z_i = \beta_0 + \beta_1 S_i + \epsilon_i, \tag{3.1}$$

여기서 $Z_i = Y_i - \overline{Y}$이고, 소지역 i, $1 \le i \le n$, 생성된 공간변수 $S_i = 1/N_i \sum_{k \in N(i)} Z_k$이다. 즉 S_i는 이웃 집합을 $N(i)$, 이웃의 수를 N_i라 했을 때 이웃으로 결정된 시군구 자료 Z_k를 평균한 값으로 구해진다. 또한 $\overline{Y} = 1/n \sum_{i=1}^{n} Y_i$이며 ϵ_i는 오차이다.

3.3. 공간정보를 활용한 이단계 소지역추정량

먼저 각 추정량 $\hat{Y}^{DE}$, $\hat{Y}^{LOGIT}$, $\hat{Y}^{LMM}$을 이용하여 Moran's I를 구한다. Moran's I 결과를 이용하여 공간상관관계가 존재한다고 판단되면 공간상관관계를 모형화한다. 일반적으로 Moran's I가 0.3 이상이면 공간상관관계를 사용하여 모형화 할 경우 우수한 결과를 얻을 수 있다고 알려져 있다. 각 통계량에 따라 다른 이웃정보시스템을 사용할 수 있으나, 본 연구에서는 같은 이웃정보시스템을 사용하였다. 표 3.1이 본 논문에서 사용된 이웃정보시스템이다. 2001년 기준으로 행정구역코드가 작성되었으며 행정구역코드 35330은 얻어진 자료가 없어 분석에서 제외하였다.

표 **4.1.** 시군구별 모집단 및 실업자 수

행정구역코드	모집단 수	실업자 수	행정구역코드	모집단 수	실업자 수
35010	4143	233	35310	483	23
35020	2043	113	35320	233	3
35030	1923	93	35340	503	3
35040	863	13	35350	243	3
35050	1213	13	35360	493	3
35060	893	43	35370	643	13
			35380	723	13

이웃정보시스템이 정해지면 이를 이용하여 공간변수를 만들 수 있다. 예를 들어 $\widehat{Y}^{LMM}$의 이단계 소지역추정법을 간단히 설명하면 다음과 같다.

(1) $\widehat{Y}^{LMM}$의 평균 $\widehat{\overline{Y}}^{LMM}$을 이용하여 $Z_i^{LMM} = \widehat{Y}_i^{LMM} - \widehat{\overline{Y}}^{LMM}$를 구한다.

(2) 이웃집합 $N(i)$에 속한 자료의 평균 $S_i^{LMM} = 1/N_i \sum_{k \in N(i)} Z_i^{LMM}$, 즉 공간변수를 구한다.

(3) Z^{LMM}를 종속변수로, S^{LMM}을 종속변수로 하는 단순회귀분석을 실시하여 예측값, $\widehat{Y}^{SP,LMM}$을 구한 후 이를 소지역추정치로 사용한다.

같은 방법으로 $\widehat{Y}^{SP}$, $\widehat{Y}^{SP,LOGIT}$을 구한다. 여기서 $\widehat{Y}^{SP}$는 $\widehat{Y}^{DE}$를 이용하여 구한다.

4. 모의실험

본 연구에서 제안한 이단계 소지역추정량의 성능을 살펴보기 위해 모의실험을 실시하였다. 모의실험에 사용된 자료는 2001년 경제활동자료 중에서 공간상관관계가 높게 나온 1,436개의 전라북도 자료로 이강석과 신기일 (2008)의 결과에 의하면 이 지역에서 Moran's I가 0.6이상으로 높게 나왔다. 다음으로 모집단 생성을 위하여 Salvati 등 (2010)의 방법을 사용하였다. 즉 1,436개의 자료를 10회의 재추출(복원허용)을 실시해 크기가 14,360개인 의사모집단(pseudo population)을 생성하였다. 또한 소지역 32320, 32340, 32350, 32360의 경우 실업자가 한명도 조사가 되지 않았기 때문에 이들 지역의 영향력을 제거하기 위하여 모든 지역에는 임으로 3개의 실업자 자료를 추가하였다. 따라서 최종적으로 생성된 모집단 수는 14,399이다. 다음으로 분석에 사용된 보조변수는 학력과 연령이다. 이는 경기도 소지역 자료 분석에서 얻어진 결과이다. 그러나 얻어진 전라북도 자료에는 보조변수인 학력과 연령이 없다. 이에 학력과 연령은 경기도 자료를 이용하여 경기도와 같은 특성을 갖도록 자료를 생성하였다. 표 4.1이 시군구별 모집단 수 및 실업자 수이다.

표 4.1의 모집단 수 및 실업자 수를 비교하면 35010, 35020 등 시인 경우의 실업률이 35310, 35320 등 군에 비하여 높은 것을 확인할 수 있다. 다음으로 본 논문에서 사용한 비교 통계량은 BIAS, MSE 그리고 MAE이며 다음과 같이 정의된다.

$$\text{BIAS} = \frac{1}{R}\frac{1}{n}\sum_{r=1}^{R}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i^{(r)}\right),$$

$$\text{MSE} = \frac{1}{R}\frac{1}{n}\sum_{r=1}^{R}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i^{(r)}\right)^2,$$

$$\text{MAE} = \frac{1}{R}\frac{1}{n}\sum_{r=1}^{R}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i^{(r)}\right|,$$

표 **4.2.** 비교 통계량 결과(전라북도)

	추정량	BIAS	MAE	MSE
일단계 추정량	$\hat{Y}^{DE}$	378.28	378.37	453904
	$\hat{Y}^{LOGIT}$	378.28	378.60	489764
	$\hat{Y}^{LMM}$	378.28	378.30	472258
이단계 추정량	$\hat{Y}^{SP}$	374.04	374.44	460420
	$\hat{Y}^{SP,LOGIT}$	318.87	350.56	459239
	$\hat{Y}^{SP,LMM}$	318.87	344.98	442207

여기서 소지역의 수 $n = 13$이고, 반복수 $R = 1{,}000$을 사용하였다. 이상을 이용하여 얻은 결과는 다음과 같다.

먼저 표 4.2의 편향(Bias)을 살펴보자. 일단계 추정량인 $\hat{Y}^{DE}$, $\hat{Y}^{LOGIT}$, 그리고 $\hat{Y}^{LMM}$의 편향은 모두 동일하다. 또한 이단계 추정량의 경우 $\hat{Y}^{SP,LOGIT}$와 $\hat{Y}^{SP,LMM}$이 같은 값을 주고 있으며 $\hat{Y}^{SP}$는 상대적으로 큰 값을 주고 있다. 여섯 개의 추정량을 편향을 기준으로 비교하면 $\hat{Y}^{SP,LOGIT}$와 $\hat{Y}^{SP,LMM}$이 가장 우수한 결과를 주고 있으며 일단계 추정량과 이단계 추정량을 비교하면 공간관계를 추가한 이단계 추정량이 우수한 결과를 주고 있음을 확인할 수 있다. MAE 비교 결과 편향과 유사한 결과를 주고 있다. 다만 $\hat{Y}^{SP,LOGIT}$보다 $\hat{Y}^{SP,LMM}$이 우수한 결과를 주고 있어 결과적으로 $\hat{Y}^{SP,LMM}$이 가장 우수하다.

다음으로 MSE를 살펴보자. 일단계 추정량을 비교하면 직접추정량인 $\hat{Y}^{DE}$가 가장 우수한 결과를 주고 있다. 이러한 결과는 직접추정량과 로지스틱회귀추정량 또는 로지스틱혼합추정량의 비교에서 흔히 있는 결과이다. 일반적으로 확실한 보조정보가 있어 이를 이용할 경우에는 $\hat{Y}^{LOGIT}$ 또는 $\hat{Y}^{LMM}$이 우수한 결과를 주는 것으로 알려져 있다. 본 논문에는 실업의 유무에 영향을 주는 보조정보로 연령과 학력이 사용되었는데 이 독립변수의 정보가 충분하지 않아 직접추정량인 $\hat{Y}^{DE}$가 가장 우수한 결과를 주고 있다.

다음으로 이단계 추정량을 살펴보자. 본 모의실험의 목적중의 하나가 이단계 추정량을 사용할 때 일단계 추정량보다 우수한 결과를 주는지 확인하는 것이다. 결과를 보면 일단계 추정량인 $\hat{Y}^{LOGIT}$과 $\hat{Y}^{LMM}$에 비해 이단계 추정량인 $\hat{Y}^{SP,LOGIT}$과 $\hat{Y}^{SP,LMM}$이 매우 우수한 결과를 주고 있음을 확인할 수 있다. 특히 $\hat{Y}^{SP,LMM}$은 모든 비교 통계량을 비교했을 때 가장 우수한 결과를 주고 있다.

5. 결론

본 논문에서는 일단계 소지역추정량으로 단위수준자료가 얻어지고, 유의미한 보조정보가 있을 경우에 흔히 사용하는 로지스틱회귀추정량과 로지스틱혼합추정량을 살펴보았다. 다음으로 일단계에서 얻어진 소지역추정량을 지역수준의 자료로 생각하고, 공간상관관계를 접목한 이단계 소지역추정량을 제안하였다. 제안된 소지역추정량은 단위수준자료가 갖고 있는 보조변수의 단위수준정보와 지역수준정보인 공간상관관계를 결합하여 추정의 정도를 향상시키는 방법이다.

모의실험에서도 확인할 수 있는 것처럼 $\hat{Y}^{SP,LMM}$, 공간로지스틱혼합추정량이 가장 우수한 결과를 주고 있다. 물론 이 추정량을 사용하기 위해서는 유의미한 보조정보가 있어야 하고 또한 공간상관관계가 있어야 한다. 그러나 향후 유의미한 보조정보가 구해지고 또한 공간상관관계 있다면 매우 유용하게 사용될 수 있을 것으로 기대된다.

참고문헌

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

신기일, 최봉호, 이상은 (2007). 공간 통계 활용에 따른 소지역 추정법의 평가, <응용통계연구>, **20**, 229–244.

여인권, 손경진, 김영원 (2008). 일반화추정방정식을 활용한 소지역 추정과 실업률 패널분석, <응용통계연구>, **21**, 665–674.

이강석, 신기일 (2008). 격자자료분석을 위한 이웃정보시스템의 비교, <응용통계연구>, **21**, 387–397.

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

최영아 (2004). 로지스틱 회귀를 활용한 소지역 실업률 추정에 관한 연구, 숙명여대 대학원, 석사 논문.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*, John Wiley and Son, New York.

Hwang, H.-J. and Shin, K.-I. (2011). Logistic regression type small area estimations based on relative error, *The Korean Journal of Applied Statistics*, **24**, 445–453.

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

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.

Two Stage Small Area Estimation

Sang Eun Lee[1] · Key-Il Shin[2]

[1]Department of Statistics, Kyunggi University
[2]Department of Statistics, Hankuk University of Foreign Studies

(Received December 30, 2011; Revised January 30, 2012; Accepted March 6, 2012)

Abstract

When Binomial data are obtained, logit and logit mixed models are commonly used for small area estimation. Those models are known to have good statistical properties through the use of unit level information; however, data should be obtained as area level in order to use area level information such as spatial correlation or auto-correlation. In this research, we suggested a new small area estimator obtained through the combination of unit level information with area level information.

Keywords: Unit level data, area level data, logit regression model, logit mixed model, spatial estimator.

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 (2009-0072926).
[2]Corresponding author: Professor, Department of Statistics, Hankuk University of Foreign Studies, Yonginsi, Kyunggi 449-791, Korea. E-mail: keyshin@hufs.ac.kr