한국통계학회논문집
2009, 16권, 2호, 265-275

오차항이 SAR(1)을 따르는 공간선형회귀모형에서 일반화 최대엔트로피 추정량에 관한 연구

전수영[1,a], 임성섭[b]

[a]고려대학교 세종캠퍼스 경제통계 산학협력단, [b]하나은행

요 약

지역적 공간의 특성을 고려한 공간선형회귀모형을 다루는 대부분의 연구들에서 사용되고 있는 자료는 완전한 상태임을 고려하고 있다. 하지만 공간선형회귀모형을 정확히 추론함에 있어서 완전한 자료가 사용가능한 경우는 그다지 많지가 않은 것이 현실이다. 만약 이러한 상황을 고려하지 않고 통계적 추론을 할 경우 잘못된 결론이 도출될 수 있다. 본 연구에서는 오차항이 일차 공간자기상관을 따르는 공간선형회귀모형에서 자료가 불완전한 상태일 경우 일반화 최대엔트로피 형식을 이용하여 미지의 모수를 추정하는 방법을 제안하였고 몬테카를로 모의실험을 통하여 여러 전통적인 추정량들과 효율성을 비교하였다. 그 결과, 자료가 불완전한 상태에서 일반화 최대엔트로피 추정량이 다른 추정방법들에 비해 효율적인 추정치를 제공하였다.

주요용어: 공간선형회귀모형, 정보 복구, GME 추정.

1. 서론

최근 생물통계학, 공간통계학 등 다양한 분야에서 방대한 자료들이 얻어지고 있으나 실제로 완벽한 자료를 얻는 경우가 매우 드물다. 또한 이러한 자료의 회귀모형을 추론함에 있어서 여러 기본 가정을 만족하는 완전한 자료가 얻어지는 경우도 그다지 많지가 않다. 따라서 자료의 생성 과정 중 어떠한 제한이 주어졌다고 할 때 우리가 지금까지 사용해 왔던 전통적인 추론의 기초를 이루는 모형은 불완전한 상태(ill-posed)의 모형이라고 할 수 있다 (송석헌과 전수영, 2006). 예를 들어 선형회귀모형에서 관찰값들을 이용해 미지의 모수를 추정하고자 할때, 만약 보통 최소제곱(Ordinary Least Squares: OLS)이나 일반화 최소제곱(Generalized Least Squares: GLS), 최대우도(Maximum Likelihood: ML)와 같은 전통적인 추정방법에 있어 설명변수와 종속변수에 포함된 정보가 충분하지 않다거나, 알려져 있는 정보(불충분한 정보)만을 이용해서 모수를 추정해야 한다면, 관측개수보다 추정해야 할 모수의 수가 더 많기 때문에 임의의 모수가 발생할 수 있고 그 결과 값이 정의되지 않을 수 있다. 또한 추정치가 매우 불안정하여 편의가 크고 정도(精度)가 낮은 추정치를 얻을 수 있다. 이와 같은 자료의 불완전한 상태는 통계학뿐만 아니라 경영 또는 경제학 분야에서 쉽게 접할 수 있는 것이다. 이러한 상황에서 모수를 추정하기 위해서는 조사자가 모수에 대한 어떠한 가정이나 제약을 하여야 한다. 그렇지 않고 전통적인 추정방법들을 사용할 경우에는 위에서 언급한 것처럼 효율적인 추정치를 얻을 수 없다.

이러한 불완전한 상태의 회귀모형에서의 추정의 문제에 대한 연구는 Golan (1994), Judge와 Golan (1992) 등이 진행하였고, 최근에 이를 바탕으로 송석헌과 전수영 (2006)이 일반화 최대엔트로피(generalized maximum entropy: GME) 방법을 패널모형에 적용하여 기존의 전통적인 추정방법들보다 가정이나 제약된 상황에 덜 민감한 강건한(robust) 새로운 추정량을 제안하였다. 이에 본 연구에

[1] 교신저자: (339-700) 충남 연기군 조치원읍 서창리 208, 고려대학교 세종캠퍼스 경제통계 산학협력단, 연구교수.
E-mail: scheon@korea.ac.kr

서는 공간자료에 대한 정보 복구(information recovery)의 효율성을 개선하는 목적으로 송석헌과 전수영 (2006)이 제안한 GME방법을 이용하여 GME추정량을 유도하고 추정량의 성질을 연구하고자 한다.

지역별로 수집된 공간자료에 대한 공간선형회귀모형(spatial linear regression model)의 분석이 전통적인 회귀분석에 비해 자료들 간의 상관관계를 고려함으로써 지역 또는 공간 특성의 변동 등 복잡한 문제에 대한 해답을 제공한다는 장점에도 불구하고 불완전한 상태의 공간자료의 문제점을 극복하려는 연구는 매우 미진한 상태이다. 최근의 여러 공간선형회귀모형연구 (Anselin과 Bera, 1998; Anselin, 2002)들도 자료가 완전하다는 가정 아래서 다양한 추정방법, 검정방법과 응용분야들을 제안하고 이에 대한 성질을 연구하였을 뿐 자료가 불완전한 상태에서의 추론의 문제는 다루지 못하였다. 따라서 본 연구에서는 관측된 공간자기상관자료가 불완전한 상태인 공간자기회귀모형에서 GME방법을 이용하여 모수들의 GME추정량을 유도하고 기존의 공간자기회귀모형에서 이용하는 다양한 추정량들(OLS, ML, GLS)과의 효율성 비교를 통하여 GME추정량의 강건한 성질에 대하여 다루고자 한다.

본 논문은 다음과 같이 구성되어 있다. 먼저 2장에서는 일반화 최대엔트로피 형식이 사용될 공간선형회귀모형에 대해 구체적으로 서술하고 공간선형회귀모형에서의 OLS, GLS, ML 추정량들을 소개한다. 그리고 3장에서는 공간선형회귀모형에 대한 GME추정량을 유도한다. 4장에서는 모의실험을 통한 각 추정량의 효율성을 비교하고, 5장에서 결론을 맺도록 한다.

2. 공간선형회귀모형

일반적으로 우리가 다루는 일반 선형회귀모형은 $y = X\beta + \epsilon$으로써 오차항 ϵ이 $E(\epsilon) = 0$, $E(\epsilon\epsilon') = \sigma^2 V$라는 조건하에 각 관측값들의 특징이나 개체들이 수집된 시간 등을 고려하여 왔다. 하지만 어떠한 자료가 지역적 특색을 알아보기 위해 지역별로 수집되었다는 공간이라는 특성을 고려해야 할 경우라면, 이것은 지역적 공간의 특성을 고려한 공간선형회귀모형을 이용하여야한다 (이재준, 2002).

어떠한 자료가 지역적 특색을 알아보기 위해 지역별로 수집되었다면 일반적으로 그러한 자료를 공간자료라고 한다. 이러한 공간자료는 특정 위치나 지역에 대한 정보가 포함되어 있어 공간 종속성(Spatial Dependence)과 공간 이분산성(Spatial Heteroskedasticity)이라는 특징을 가질 수 있다. 먼저 공간 종속성은 공간 자기상관(spatial autocorrelation)이라고도 부르는데, 이것은 모집단 구성원이 지리적 위치를 통하여 관계되었을 경우 발생한다. 여기에서의 자기상관은 회귀분석이나 시계열 자료분석에서 거론되는 자기상관에서 도출된 개념이지만, 여러 중요한 차이점을 가진다. 즉, 시간은 한 방향에서의 문제를 다루지만 공간은 한 가지 이상의 방향을 가질 수 있다는 점, 시간은 순서화 된 자료를 제공하나 공간에서는 한 지역에서의 오차가 다른 지역에 영향을 주며 순서화 된 자료를 제공하지는 않는다는 점, 시계열 자료는 일반적으로 등간격으로 측정되지만 공간 자료는 불규칙한 격자에서도 관측될 수 있다 (Dubin, 1998)는 점 등 여러 중요한 다른면을 공간자료는 가진다. 공간 효과의 두 번째 형태인 공간 이분산성은 모수가 지리적 위치에 따라 변하며 자료가 등분산이 아닌 경우인 안정성의 결핍과 관련된다. 예를 들어 어떠한 부유한 지역과 가난한 지역과 같은 유사하지 않은 공간에서의 횡단면 자료를 통한 추정에서 이러한 문제가 발생할 수 있다. 하지만 이러한 이분산성의 문제는 대부분의 일반적인 통계기법에 의해 해결될 수 있다.

따라서, 공간선형회귀모형에서 공간자료의 여러 특징 중 공간 자기상관을 고려할 필요가 있으며, 이에 본 논문에서는 오차항이 일차 공간 자기상관(Spatial First-order Autoregressive: SAR (1)) 관계가 있는 경우를 고려한다. 이러한 경우에 어떻게 공간 자기상관의 구조를 공간선형회귀모형에 표현할 것인가 하는 문제가 발생한다. 이러한 문제는 공간상에서 서로 다른 위치 또는 지역들 간의 상호작용을 어떻게 수식화 할 것인가로 나타내어질 수 있다. 이 문제는 Tobler (1970)의 지리학의 제 1 법칙인 "근접한 관찰개체일수록 오차항이 서로 밀접하게 관련되어 있고 거리가 멀어질수록 이 관계는 감소한

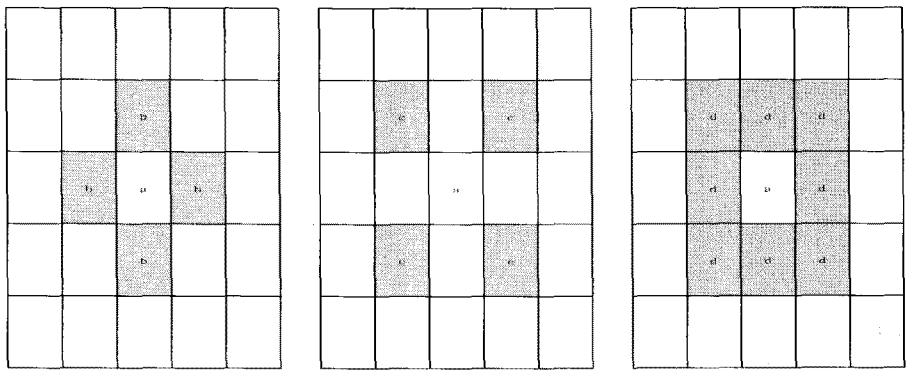

그림 1: *인접 유형 ((a) Rook형 (b) Bishop형 (c) Queen형)*

다."는 개념을 통해 정의될 수 있다. 공간통계학에서는 이러한 인접 정보들을 공간 가중행렬로 표현한다.

보통 T개의 공간자료에서 관측치 i는 최대 $(T-1)$개의 인접 관측치를 가질 수 있다. 예를 들어 공간 가중행렬의 가장 단순한 형태인 이진인접행렬(binary contiguity matrix)은 다음과 같다.

$$w_{ij} = \begin{cases} 1, & \text{만일 } i\text{와 } j\text{가 이웃하면,} \\ 0, & \text{만일 } i\text{와 } j\text{가 이웃하지 않으면.} \end{cases} \tag{2.1}$$

이진인접행렬과는 달리 일반화 가중행렬은 조사자가 사전의 고려에 의해 적당하다고 생각되는 가중치를 선택할 수 있다. 따라서 이것은 다양한 형태의 가중행렬을 정의할 수 있다는 장점이 있다. Cliff와 Ord (1973)는 두 지역 간의 거리와 둘레 길이를 이용하여 인접행렬보다는 좀 더 확장된 가중행렬을 제안하기도 하였다. 보통 인접 유형은 3가지로 나눌 수 있는데 첫째는 두 인접 관측치 a와 b가 공통변을 공유하는 경우이고, 다음은 두 인접 관측치 a와 c가 꼭지점을 공유하는 경우이다. 마지막으로 두 인접 관측치 a와 d가 꼭지점이나 공통변을 모두 공유하는 경우로, 위의 세 가지 경우를 체스 게임에서의 행마법에 따라 Rook형, Bishop형, Queen형으로 부를 수 있다 (그림 1).

하지만 이러한 공간 가중행렬은 독립적으로 추정하는 것이 불가능하기 때문에 사전에 미리 지정되어야 하는 것으로 간주한다. 또한 일반적으로 해석과 분석의 용이함, 특히 정상성 조건을 만족시키기 위하여 가중행렬의 각 행의 합을 1로 만드는 행 표준화가 필요하고 (Kelejian과 Prucha, 1998, 1999)이 행렬을 대칭행렬로 간주하며 대각원소는 0으로 놓는다.

보통 일반적인 공간선형회귀모형은 다음과 같이 표현할 수 있다.

$$y = \lambda W_1 y + X\beta + \epsilon, \quad \epsilon = \rho W_2 \epsilon + u. \tag{2.2}$$

여기서 y는 $T \times 1$ 종속변수 벡터이고 X는 $T \times K$ 계획행렬로 비확률적인 설명변수 행렬이다. 그리고 β는 $K \times 1$ 모수 벡터이고 λ는 공간시차변수의 자기상관계수이며 ρ는 오차항 ϵ에 대한 자기상관계수이다. W_1과 W_2는 $T \times T$ 공간가중행렬이고 u는 $E(u) = 0$이고 $E(uu') = \sigma_u{}^2\Omega$인 정규분포를 따르는 $T \times 1$ 오차항 벡터이다.

식 (2.2)에서 오차항이 일차 공간자기회귀과정을 따르는 경우는 $\lambda = 0$인 경우로 공간선형회귀모형을 다음과 같이 나타낼 수 있다.

$$y = X\beta + (I - \rho W)^{-1}u, \quad u \sim N\left(0, \sigma_u{}^2\Omega\right). \tag{2.3}$$

여기서 I는 T차원의 단위행렬이다. 또한 y의 분산-공분산 행렬을 좀 더 정확하게 표현하면 다음과 같다.

$$E(yy') = \sigma_u{}^2\Omega^*. \tag{2.4}$$

여기서 $\Omega^* = (I - \rho W)^{-1}\Omega(I - \rho W')^{-1}$이며 $(I - \rho W)$는 정칙행렬이다.

2.1. OLS, GLS 추정량

오차항이 일차 공간자기회귀과정을 따르는 모형 (2.3)에서 회귀계수 β에 대한 OLS추정량은 다음과 같이 구해진다.

$$\widehat{\beta}_{OLS} = (X'X)^{-1}X'y. \tag{2.5}$$

OLS추정량을 볼 때, 이 추정량이 오차항에 대해 등분산을 고려하기 때문에 SAR(1) 모형에서 불편추정량이 될 수는 있지만 다른 추정량에 비해 효율성이 떨어지게 될 것이다. 또한 OLS 추정에 의하여 추정된 표준오차는 편향된다는 사실이 잘 알려져 있다 (Moulton, 1986). OLS추정량에 의한 잔차는 $\widehat{u}_{OLS} = y - X\widehat{\beta}_{OLS}$이다.

만일 식 (2.4)에서 분산성분들이 알려져있다면, β에 대한 GLS추정량은 다음과 같이 구해진다.

$$\tilde{\beta}_{GLS} = \left(X'\Omega^{*-1}X\right)^{-1}X'\Omega^{*-1}y. \tag{2.6}$$

GLS추정량은 OLS추정량과는 달리 오차항에 대해 이분산의 경우도 고려하고 있기 때문에 OLS추정량보다 효율성이 좋을 것이다. 그러나 현실적으로 분산성분들은 알려지지 않은 경우가 대부분이므로 GLS추정량은 단지 이론적인 추정량에 불과하지만 다른 추정량과의 비교시 기준이 되는 추정량으로 사용될 수 있다.

2.2. ML 추정량

먼저 ϵ에 대한 우도함수 L은 다음과 같다.

$$L = \left(2\pi\sigma_u{}^2\right)^{-\frac{T}{2}}|\Omega^*|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\sigma_u{}^2}(y - X\beta)'\Omega^{*-1}(y - X\beta)\right\}|I - \rho W|. \tag{2.7}$$

위 식의 양변에 log를 취하여 β와 $\sigma_u{}^2, \rho$에 대해 각각 편미분을 통해 다음의 $\hat{\beta}$을 구할 수 있다.

$$\begin{aligned}\hat{\beta}_{ML} &= (X'(\Omega^{**})^{-1}X)^{-1}X'(\Omega^{**})^{-1}y, \\ \Omega^{**} &= (I - \hat{\rho}W)^{-1}\Omega(I - \hat{\rho}W')^{-1}.\end{aligned} \tag{2.8}$$

ML추정량은 GLS추정량과 같지만 그 차이는 ρ의 추정 유무와 관련된다. 즉 식 (2.6)과 (2.8)에서 보는 바와 같이 GLS추정량에서와는 달리 ML추정량에서는 추정치 $\hat{\rho}$을 사용하여 β를 추정하게 된다 (Griffith, 1988). 따라서 ρ와 $\sigma_u{}^2$이 알려져 있을 경우에는 $\hat{\beta}_{GLS}$와 $\hat{\beta}_{ML}$이 동일하게 된다.

3. 회귀계수에 대한 GME 추정량

이 장에서는 식 (2.3)에 대한 GME추정량을 구하기 위해 회귀계수와 오차항을 재모수화 하는 과정을 알아본다. GME추정량을 구하는데 이용되는 최대엔트로피방법에 대한 자세한 내용은 송석헌과 전수영 (2006)을 참조하길 바란다.

식 (2.3)에서 알려지지 않은 모수 $\beta = (\beta_1, \ldots, \beta_k)$와 오차 u가 제한된 사전정보를 가졌다고 가정한다. 예를 들어 모수 β와 오차 u가 임의의 범위를 가지고 있다고 가정을 하면 모수 β와 오차 u가 사전 가중치나 유한 대응치(finite supports)를 가지는 이산확률변수로 구성되어 있다고 할 수 있다. 따라서 다음과 같이 모수 β와 오차 u를 재모수화(reparameterization) 할 수 있다. 각 β_k를 대응치 z와 $M(2 \le M < \infty)$개의 가능한 결과치로 구성된 이산확률변수라고 가정한다. 만일 z_{k1}과 z_{kM}이 각각 모수 β_k의 가능한 극한값들(상한값과 하한값)이라고 하면 β_k를 이러한 두 점의 볼록 조합(convex combination)으로 표현할 수 있다. 즉 다음과 같이 $M = 2$에 대해 각 극한값에 확률을 할당한다.

$$\beta_k = p_k z_{k1} + (1 - p_k) z_{kM}, \quad k = 1, \ldots, K. \tag{3.1}$$

여기서 모수공간은 $\pounds = (\beta_1, \ldots, \beta_K) \subset \mathcal{R}^K$에 의해 표현되며 $p_k \in [0, 1]$이다.

모수들을 일반적으로 표현하기 위하여 z_k를 £의 k번째 차원까지 확장된 M개의 점들의 집합이라고 하자. $\sum_{k=1}^{K} p_k = 1$인 양의 가중치가 주어졌을 때 k번째 모수는 가중치 p_k를 가지는 점 z_k의 볼록 조합으로서 표현된다. 더 나아가 이러한 볼록 조합은 어떠한 $\beta \in \text{int}(\pounds)$라도 행렬로써 표현할 수 있으므로 다음과 같이 회귀계수 β를 재모수화 할 수 있다.

$$\beta = Zp = \begin{bmatrix} z_1' & 0 & \cdots & 0 \\ 0 & z_2' & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & z_K' \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_K \end{bmatrix}. \tag{3.2}$$

여기서 $\beta_k = \sum_{m=1}^{M} z_{km} p_{km}$, $k = 1, 2, \ldots, K$, $z_k = (z_{k1}, z_{k2}, \ldots, z_{kM})'$, $p_k = (p_{k1}, p_{k2}, \ldots, p_{kM})'$이다.

다음으로 오차항 u를 유한한 모수들을 가지는 확률벡터라고 가정한다. 각 u_t를 $J(2 \le J < \infty)$개의 가능한 결과치를 가지는 이산확률변수라고 하고, 만일 v_{t1}과 v_{tJ}가 각각 오차 u_t의 가능한 극한값들(상한값과 하한값)이라고 하면, u_t를 이러한 두 점의 볼록 조합으로 표현할 수 있다.

$$u_t = w_t v_{t1} + (1 - w_t) v_{tJ}, \quad t = 1, \ldots, T. \tag{3.3}$$

여기서 $w_t \in [0, 1]$이다. 오차항 또한 행렬의 형태로 재모수화하면 다음과 같다.

$$u = Vw = \begin{bmatrix} v_1' & 0 & \cdots & 0 \\ 0 & v_2' & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & v_T' \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_T \end{bmatrix}. \tag{3.4}$$

여기서 $u_t = \sum_{j=1}^{J} v_{tj} w_{tj}$, $t = 1, 2, \ldots, T$, $v_t = (v_{t1}, v_{t2}, \ldots, v_{tJ})'$, $w_t = (w_{t1}, w_{t2}, \ldots, w_{tJ})'$이다. 다음으로 일반화 최대엔트로피(GME) 형식을 이용하여 모수에 대한 GME추정량을 구한다.

$$\max_{p,w} H(p, w) = -p' \ln p - w' \ln w. \tag{3.5}$$

첫 번째 조건은 일치성으로 다음과 같다.

$$y = XZp + (I - \rho W)^{-1}Vw. \tag{3.6}$$

여기서 $y_t = \sum_{k=1}^{K} x_{kt}(\sum_{m=1}^{M} z_{km}p_{km}) + \sum_{k=1}^{T} a_{tk}(\sum_{j=1}^{J} v_{kj}w_{kj})$, $t = 1, 2, \ldots, T$이고, a_{tk}는 $(I - \rho W)^{-1}$의 (t, k)번째 원소이다. 두 번째 조건은 정규화 가법성으로 다음과 같다.

$$i_K = (I_K \otimes i_M')p, \quad i_T = (I_T \otimes i_J')w. \tag{3.7}$$

여기서 $\otimes$는 크로네커곱(Kronecker product)을 나타내며, I_K는 K차원의 단위행렬이고, i_T는 모든 원소가 1인 크기가 T인 벡터이다. 이제 GME추정량 $\tilde{\beta}_{GME}$를 유도해 보자. 먼저 라그랑지안 방정식을 다음과 같이 정의한다.

$$\begin{aligned} L \;=\; & -p' \ln p - w' \ln w + \lambda'[y - XZp - (I - \rho W)^{-1}Vw] + \\ & \theta'[i_K - (I_K \otimes i_M')p] + \tau'[i_T - (I_T \otimes i_J')w]. \end{aligned} \tag{3.8}$$

식 (3.8)에서 $p, w, \lambda, \theta, \tau$ 각각의 값에 대해 1차 편미분을 하여 $\tilde{p}$ 추정량을 구해보면 다음과 같다.

$$\tilde{p} = \exp(-Z'X'\lambda) \odot \{(I_K \otimes i_M i_M') \exp(-Z'X'\lambda)\}^{-1}. \tag{3.9}$$

여기서 $\odot$는 하다마드곱(Hadamard product)을 나타낸다. 이를 통해 회귀계수 β_{GME}를 추정하면 다음과 같다.

$$\tilde{\beta}_{GME} = Z\tilde{p} = Z \cdot \exp(-Z'X'\lambda) \odot \{(I_K \otimes i_M i_M') \exp(-Z'X'\lambda)\}^{-1}. \tag{3.10}$$

식 (3.10)에서 k번째 $\tilde{\beta}_{GME}$를 구해보면 다음과 같다.

$$\tilde{\beta}_{GME(k)} = \sum_{m=1}^{M} z_{km}\tilde{p}_{km} = \sum_{m=1}^{M} z_{km} \left[\frac{\exp\left(-z_{km} \sum_{t=1}^{T} x_{kt}\lambda \right)}{\Omega_k(\tilde{\lambda})} \right]. \tag{3.11}$$

여기서 $\Omega_k(\tilde{\lambda}) = \sum_{m=1}^{M} \exp\left[-z_{km}\left(\sum_{t=1}^{T} x_{kt}\lambda \right) \right]$이다.

4. 모의실험

4.1. 모의실험 방법

지금까지 2장과 3장에서 살펴보았던 각 추정량의 불완전 상태의 모형에서 효율성을 비교하기 위해 본 장에서는 모의실험을 통하여 그들의 효율성을 비교하고자 한다. 모의실험에 사용된 모형은 다음과 같은 오차항이 SAR(1)을 따르는 공간선형회귀모형이다.

$$y_t = \beta x_t + (I - \rho W)_t^{-1} u_t, \quad t = 1, 2, \ldots, T. \tag{4.1}$$

여기서 y는 $T \times 1$인 종속변수 벡터이며 X는 $(T \times 4)$인 독립변수 벡터이고, β는 (4×1)인 회귀계수 벡터, W는 $(T \times T)$인 가중행렬, u는 $(T \times 1)$인 오차벡터로서 정규분포 $N(0, 1)$에서 발생시켰다. $T = (10, 50, 100)$이다. 위 모형 (4.1)를 불완전한 상태에서 모의실험을 하기 위해 계획행렬 X를 다중공선성

이 존재하는 행렬로 만든다. 이를 위해 Belsley (1991)가 언급한 비정칙 함수를 사용하는데 그가 사용한 조건수 μ는 다음과 같다.

$$k(X'X) = \mu = \frac{\pi_{(1)}}{\pi_{(K)}}. \tag{4.2}$$

이것은 행렬 X의 비정칙치(singular value)의 최대값과 최소값의 비율을 나타낸 것으로 만약 계획행렬 X가 직교행렬이고 X의 열이 선형 독립이라면 $\pi_i = 1$이고 따라서 $k(X'X) = 1$이 된다. 하지만 공선성(collinearity)의 정도가 증가함에 따라 $\pi_{(K)} \to 0$이고 $k(X'X) \to \infty$가 된다. 여기서 원하는 조건수 $k(X'X) = \mu$를 가지는 행렬를 만들기 위해 행렬 X에 대해 다음의 비정칙치 분해(singular value decomposition)를 사용한다.

$$X = U\Lambda V'. \tag{4.3}$$

여기서 $\Lambda = \text{diag}(\pi_{(i)})$이고 $\pi_{(i)}$는 X의 비정칙치(singular value)이다. Belsley (1991)의 조건수 μ를 가지는 행렬 X_a를 만들기 위해 X의 비정칙치로 구성된 행렬 Λ의 대각원소를 다음과 같이 변형한다.

$$a = \left[\sqrt{\frac{2}{1+\mu}},\ 1,\ 1,\ \sqrt{\frac{2\mu}{1+\mu}} \right]. \tag{4.4}$$

따라서, 모의실험을 위해 다중공선성이 존재하는 조건수 μ를 가지는 행렬 X_a를 다음과 같이 표현할 수 있다.

$$X_a = U\Lambda_a V'. \tag{4.5}$$

본 모의실험에서 조건수 μ는 다중공선성이 존재하지 않는 1부터 공선성의 정도가 높아지는, 즉 자료의 왜곡 상태가 심해지는 $\mu = (1, 5, 10, 50, 100)$으로 정하였다. 표본의 크기 T는 $T = (10, 50, 100)$으로 하였으며, ${\sigma_u}^2$과 ρ는 기지(旣知)라고 가정한다. 먼저 ${\sigma_u}^2$은 2와 15의 두 가지 경우로 하였고, 자기상관계수 ρ의 값은 $\rho = (0, 0.2, 0.5, 0.8)$로 하였다. 또한 β와 u에 대한 compact support space z_k와 v_t는 각각 다음과 같이 나타내었다. $z_k = (-10, -5, 0, 5, 10)$이고, ${\sigma_u}^2 = 2$인 경우 $v_t = (-5, -2.5, 0, 2.5, 5)$로 하였으며 ${\sigma_u}^2 = 15$인 경우에는 $v_t = (-12, -6, 0, 6, 12)$로 하였다. 그리고 공간가중행렬 W는 Bishop형의 인접 유형을 이용한 이진인접 행렬을 사용하였으며, 행 표준화를 만족하는 단순한 형태로 w_{ij}는 $w_{1,T} = w_{2,T-1} = \cdots = w_{i,T-i+1} = 1$이고 그 외의 경우에는 $w_{i,j} = 0$으로 하였다. 여기서 $i = 1, \ldots, T$이고 $j = 1, \cdots, T$이다. 각각의 경우에 대해 1,000회의 몬테카를로 모의실험을 시행하였다.

4.2. 모의실험의 결과

표 1, 2와 3은 관측개수 T와 오차항의 분산 ${\sigma_u}^2$, 그리고 자기상관계수 ρ와 조건수 μ에 따른 각 추정치의 MSE의 모의실험 결과를 나타낸 것이다. 본 모의실험은 ρ와 ${\sigma_u}^2$이 알려져 있다는 가정하에 진행되었기 때문에 GLS추정치는 ML과 동일한 MSE를 가진다. 그래서 본 모의실험의 결과에서는 GLS추정치의 결과를 생략하였다. 모의실험의 결과를 살펴보면 먼저 조건수 μ가 1인 경우, 즉 계획행렬 X가 다중 공선성이 존재하지 않고 열들이 선형 독립을 이루는 경우에는 거의 대부분 오차항의 분산 ${\sigma_u}^2$과 자기상관계수 ρ에 상관없이 β에 대한 GME추정치의 MSE값이 OLS나 ML추정치에 비해 큰 것을 알 수 있다. 즉 GME추정치가 다른 추정치에 비해 효율성이 떨어짐을 알 수 있다. 하지만 조건수 μ를 증가시켜 가며 다중공선성의 정도를 높여감에 따라 OLS나 ML추정치의 MSE값은 커지지만 GME추정치의 MSE값은 큰 변화 없이 일정한 값을 유지함을 알 수 있다. 특히 오차항의 분산 ${\sigma_u}^2$값이 2에서 15로 커짐에 따라 조건수 μ에 따른 OLS, ML추정치는 MSE값의 변화가 매우 심하나 GME추정치는 ${\sigma_u}^2$의 변화와 무관하게 일정한 값을 유지함을 알 수 있다. 예를 들어 표 3을 보면, 먼저 조건수 μ가 1인 경우 ${\sigma_u}^2$이나 ρ에 상관없이 GME추정치의 MSE값이 OLS나 ML추정치에 비해 큼을 알 수 있다. 하

표 1: 각 추정치의 MSE (T = 10인 경우)

σ^2_u	ρ	μ	OLS	ML	GME	σ^2_u	ρ	μ	OLS	ML	GME
2	0.0	1	0.435	0.435	2.158	15	0.0	1	2.972	2.972	2.199
		5	4.571	4.571	2.457			5	33.777	33.777	2.469
		10	7.495	7.495	2.462			10	53.456	53.456	2.467
		50	27.244	27.244	2.464			50	197.718	197.718	2.474
		100	51.895	51.895	2.459			100	384.986	384.986	2.473
	0.2	1	0.433	0.391	2.141		0.2	1	3.333	3.111	2.207
		5	5.040	4.763	2.459			5	37.642	34.949	2.474
		10	7.542	6.877	2.459			10	59.282	53.805	2.492
		50	29.776	27.254	2.460			50	226.394	204.505	2.482
		100	56.499	52.095	2.464			100	451.562	424.779	2.489
	0.5	1	0.880	0.458	2.158		0.5	1	6.448	3.969	2.397
		5	9.738	5.594	2.465			5	79.575	43.597	2.507
		10	15.123	8.605	2.462			10	118.099	63.220	2.508
		50	61.692	32.469	2.465			50	461.614	245.622	2.507
		100	109.671	65.425	2.457			100	842.492	493.283	2.515
	0.8	1	4.671	0.764	2.355		0.8	1	37.766	5.824	4.500
		5	58.285	8.818	2.567			5	463.802	69.111	5.412
		10	98.837	13.616	2.548			10	642.798	94.096	5.493
		50	345.966	46.747	2.597			50	2374.621	397.252	5.624
		100	691.931	105.776	2.555			100	4999.773	819.748	5.691

표 2: 각 추정치의 MSE (T = 50인 경우)

σ^2_u	ρ	μ	OLS	ML	GME	σ^2_u	ρ	μ	OLS	ML	GME
2	0.0	1	0.044	0.044	2.048	15	0.0	1	0.335	0.335	2.038
		5	4.705	4.705	2.491			5	34.685	34.685	2.497
		10	7.250	7.250	2.493			10	53.486	53.486	2.492
		50	28.068	28.068	2.488			50	195.490	195.490	2.499
		100	54.731	54.731	2.492			100	400.466	400.466	2.493
	0.2	1	0.050	0.044	2.006		0.2	1	0.374	0.316	2.035
		5	5.198	4.514	2.492			5	36.232	32.347	2.493
		10	7.751	6.712	2.492			10	60.680	51.604	2.493
		50	29.763	24.885	2.491			50	233.521	198.953	2.497
		100	55.908	44.439	2.491			100	440.554	381.606	2.494
	0.5	1	0.099	0.039	2.027		0.5	1	0.729	0.279	2.109
		5	9.874	3.974	2.490			5	80.435	30.430	2.491
		10	15.654	6.353	2.496			10	118.542	46.410	2.502
		50	62.136	25.319	2.490			50	470.182	177.339	2.499
		100	114.922	40.324	2.491			100	839.447	331.572	2.499
	0.8	1	0.549	0.031	2.080		0.8	1	4.341	0.231	2.755
		5	60.129	3.093	2.506			5	451.124	24.308	3.060
		10	82.914	4.674	2.494			10	662.454	35.387	3.019
		50	357.091	18.240	2.508			50	2457.649	136.604	2.857
		100	704.860	35.745	2.495			100	4902.092	261.348	3.021

지만 조건수 μ가 증가할수록 OLS, ML추정치의 MSE값은 증가하지만 GME추정치의 MSE값은 큰 변화 없이 2.492 ~ 2.564 값의 매우 안정된 MSE변화량을 나타내고 있다. 더욱이 표 1을 보면 $\sigma_u{}^2$의 값이 2에서 15로 변함에 따라 조건수 μ가 증가할 때 $\sigma_u{}^2$ = 2일 때 보다 $\sigma_u{}^2$ = 15일 때 OLS, ML추정치의 MSE값의 변화의 폭이 더 크다는 것을 알 수 있다. 하지만 GME추정치의 경우에는 이러한 $\sigma_u{}^2$의 변화에 상관없이 2.141 ~ 5.691 값을 유지함을 알 수 있다.

표 3: 각 추정치의 MSE (T = 100인 경우)

$\sigma^2{}_u$	ρ	μ	OLS	ML	GME	$\sigma^2{}_u$	ρ	μ	OLS	ML	GME
2	0.0	1	0.021	0.021	1.904	15	0.0	1	0.160	0.160	1.968
		5	4.390	4.390	2.495			5	31.862	31.862	2.496
		10	6.473	6.473	2.498			10	53.961	53.961	2.499
		50	26.173	26.173	2.495			50	186.366	186.366	2.498
		100	55.513	55.513	2.496			100	374.108	374.108	2.493
	0.2	1	0.024	0.020	1.897		0.2	1	0.183	0.162	1.981
		5	5.354	4.571	2.494			5	40.330	35.007	2.496
		10	7.718	6.659	2.497			10	63.830	54.127	2.499
		50	27.687	22.980	2.495			50	243.949	214.416	2.497
		100	58.011	50.184	2.497			100	431.815	388.466	2.492
	0.5	1	0.047	0.019	1.990		0.5	1	0.373	0.137	1.998
		5	11.051	3.989	2.495			5	71.170	27.557	2.494
		10	17.383	5.680	2.497			10	108.247	39.938	2.495
		50	56.457	23.015	2.498			50	425.770	149.509	2.506
		100	118.958	46.108	2.500			100	876.424	317.512	2.495
	0.8	1	0.255	0.014	1.999		0.8	1	1.943	0.102	2.314
		5	60.755	2.943	2.500			5	438.472	20.598	2.518
		10	90.384	4.350	2.502			10	688.157	31.596	2.547
		50	329.458	16.025	2.497			50	2159.903	125.092	2.564
		100	660.010	37.728	2.500			100	4404.497	268.384	2.551

결론적으로 GME추정량은 그 추정량의 특성에 따라 완전한 상태의 자료에서는 전통적인 추정량에 비해 좋은 추정치를 제공하지 못하고 효율성이 떨어진다. 하지만 자료가 불완전한 경우에는 다른 추정량에 비해 자료의 상태에 덜 민감하여 강건한 추정치를 제공한다.

5. 결론

공간선형회귀모형에 대한 통계적 추론 문제에서 수집된 자료가 완전한 상태인지 여부는 매우 중요하다. 실제로 수집된 자료들이 비실험자료인지 역함수의 해가 존재하지 않는지 등의 불완전한 상태를 고려하지 않고 전통적인 추정방법을 사용할 경우에는, 서론에서 언급한 것처럼 만약 그 자료가 불완전 상태의 자료라면 지역 또는 공간 특성의 변동과 같은 통계적 추론을 올바르게 할 수 없게 된다. 그러므로 공간선형회귀모형에서 자료가 불완전한 상태일 때 적절한 통계적 추정을 위하여 본 연구에서 제안한 GME추정방법을 고려해 보는 것이 필요하다.

본 논문에서는 오차항이 SAR(1) 모형을 따르는 공간선형회귀모형에서의 모수 추정에 있어서 GME추정량을 유도하고 그 추정량에 대한 효율성을 모의실험을 통하여 OLS, ML, GLS추정량과 비교하여 보았다. 그 결과 자료의 불완전 상태가 심해질수록 OLS, ML, GLS추정량보다 GME추정량이 상당히 안정적이고 매우 효율적인 추정치를 제공해 주고 있다. 결론적으로, 수집된 자료의 왜곡 상태가 심한 불완전한 상태의 자료일 경우에는 전통적인 통계기법을 사용한 추정보다는 자료의 상태나 가정 등에 덜 민감한 GME추정량을 사용하는 것이 적절할 것으로 생각된다.

끝으로 본 연구는 단지 분산성분들이 알려져 있는 경우를 다루었다. 하지만 현실적으로 분산성분들이 알려져 있지 않은 경우가 대부분이므로 GLS나 ML추정량들은 단지 이론적인 추정량에 불과할 수 있다. 그래서 향후에 분산성분을 추정한 뒤 이를 이용한 회귀계수의 FGLS(Feasible GLS)추정량의 개발이 필요할 것으로 생각된다. 앞으로 이와 같은 추정량이 개발이 된다면 연구에 많은 도움이 될 것으로 생각한다.

참고 문헌

송석헌, 전수영 (2006). 패널회귀모형에서 최대엔트로피 추정량에 관한 연구, <응용통계연구>, **19**, 521–534.

이재준 (2002). 오차항이 공간자기상관을 갖는 선형회귀모형에서 회귀계수 검정에 관한 연구, <고려대학교 석사학위 논문>.

Anselin, L. and Bera, A. (1998). *Spatial Dependence in Linear Regression Models with an Introduction to Spatial Econometrics,* Handbook of Applied Economic Statistics, New York.

Anselin, L. (2002). Under the hood: Issues in the specification and interpretation of spatial regression models, *Agricultural Economics,* **27**, 247–267.

Belsley, D. (1991). *Conditioning Diagnostics: Collinearity and Weak Data in Regression,* John Wiley & Sons, New York.

Cliff, A. D. and Ord, J. K. (1973). *Spatial Autocorrelation,* Pion, London.

Dubin, R. A. (1998). Spatial autocorrelation: A primer, *Journal of Housing Economics,* **7**, 304–327.

Golan, A. (1994). A multi-variable stochastic theory of size distribution of firms with empirical evidence, *Advances in Econometrics,* **10**, 1–46.

Griffith, D. A. (1988). *Advanced Spatial Statistics: Special Topics in the Exploration of Quantitative Spatial Data Series,* Kluwer Academic Publishers.

Judge, G. G. and Golan, A. (1992). Recovering information in the case of ill-posed inverse problems with noise, *Unpublished paper,* University of California at Berkeley.

Kelejian, H. H. and Prucha, I. R. (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances, *The Journal of Real Estate Finance and Economics,* **17**, 99–121.

Kelejian, H. H. and Prucha, I. R. (1999). A generalized moments estimator for the autoregressive parameter in a spatial model, *International Economic Review,* **40**, 509–533.

Moulton, B. R. (1986). Random group effects and the precision of regression estimates, *Journal of Econometrics,* **32**, 385–397.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region, *Economic Geography,* **46**, 230–240.

2008년 12월 접수; 2009년 1월 채택

Generalized Maximum Entropy Estimator for the Linear Regression Model with a Spatial Autoregressive Disturbance

Sooyoung Cheon[1,a], Seong-Seop Lim[a]

[a]KU Industry-Academy Cooperation Group Team of Economics and Statistics, Korea Univ.,
[a]Personal Risk Management Team, Hana Bank

Abstract

This paper considers a linear regression model with a spatial autoregressive disturbance with ill-posed data and proposes the generalized maximum entropy(GME) estimator of regression coefficients. The performance of this estimator is investigated via Monte Carlo experiments. The results show that the GME estimator provides efficient and robust estimate for the unknown parameter.

Keywords: Spatial linear regression model, information recovery, GME estimation.

[1] Corresponding author: Research Professor, KU Industry-Academy Cooperation Group Team of Economics and Statistics, Korea University, Jochiwon 339-700, Korea. E-mail: scheon@korea.ac.kr