응용통계연구 제 19권 3호, 2006년, pp. 521–534

패널회귀모형에서 최대엔트로피 추정량에 관한 연구*

송석헌[1] 전수영[2]

요 약

횡단면 자료와 시계열 자료가 병합된 패널회귀모형을 다루는 대부분의 연구들에서 사용되고 있는 자료는 완전한 자료를 고려하고 있다. 그러나, 실제적으로 완전한 자료보다는 불완전한 자료가 많다. 이러한 상황을 고려하지 않고 통계적인 추론을 하게 되면 잘못된 결론이 도출될 수 있다. 따라서, 자료의 형태를 충분히 고려한 추정량을 바탕으로 자료를 분석해야 한다. 본 연구는 패널회귀모형에서 자료가 불완전 상태인 경우 최대엔트로피 형식을 이용한 일반화최대엔트로피 추정량을 제안하고, 추정량들의 효율성을 모의실험을 통하여 비교하였다. 모의실험 결과, 일반화최대엔트로피 추정량이 가장 안정적이고 효율적인 추정량임을 보여주었다.

주요용어: 패널회귀모형, 정보복구, ME추정, GME추정

1. 서론

최근 생물정보학, 의학, 환경학 등 여러 분야에서 방대하고 다양한 자료가 얻어지고 있으나, 실제 얻어지는 자료가 완벽한 경우가 드문 것이 사실이다. 즉, 자료의 생성과정에 제한이 주어졌다고 할 때, 자료가 불완전 상태(ill-posed)라고 한다. 따라서 통계적으로 올바른 모형의 설정과 모형에 대한 적절한 통계적 추론을 위해서는 미비한 자료의 복구 또는 정보 복구(information recovery)의 중요성이 강조되고 있다. 예를 들어 회귀모형에서 불완전 상태(ill-posed)가 되는 양상은 다음과 같은 이유 때문에 흔히 일어난다. i) 평균이 추세를 가지는 비정상성(nonstationarity)이거나, ii) 하나의 모형을 설정할 때 그 모형에 2개 또는 그 이상의 올바르지 않은 원소가 포함되어 실제와 다른 모형이 설정되어져서 미지 모수의 수가 관찰된 자료의 수를 초과하는 경우, 또는 iii) 설명변수가 공선성인 경우에도 흔히 나타난다. 따라서 만일 회귀모형에서 설명변수와 종속변수 안에 포함된 정보가 충분치 않거나, 또는 알려져 있는 불충분한 정보만을 이용하여 전통적인 추정방법으로 회귀계수를 추정을 한다면 추정량의 해가 정의되지 않을 수도 있으며, 설령 구한다 하여도 추정량의 분산의 편의가 크고, 정도가 낮게 될 것이다.

* 본 연구는 KRF(2004-C00073-101102)연구비 지원에 의하여 수행되었음.
1) (136-701) 서울시 성북구 안암동 5-1, 고려대학교 통계학과, 교수
E-mail: ssong@korea.ac.kr
2) (136-701) 서울시 성북구 안암동 5-1, 고려대학교 통계학과
E-mail: sycheon@korea.ac.kr

이러한 불완전 상태의 회귀모형에서의 추정의 문제에 대한 연구는 Golan(1994), Judge and Golan (1992) 등이 진행하였으나 커다란 진척이 없었다. 그러므로 불완전한 상태의 자료 처리문제와 함께 정보 복구의 효율성을 증가시키는 추정량을 구해야 하는 필요성이 대두되고 있다.

이에 본 연구에서는 정보 복구의 효율성을 개선하는 목적으로 최대엔트로피(maximum entropy, ME) 방법을 이용하고자 한다. 특히 이를 확장한 일반화 최대엔트로피(generalized maximum entropy, GME)방법을 적용하여 새로운 추정량을 유도하고 추정량의 성질을 연구하고자 한다. 최근 Golan et al. (1994, 1996)) 등이 불완전 상태(ill-posed)의 자료 문제에 초점을 두고 이를 극복 할 수 있는 해결방안으로 GME방법을 적용하는 연구하였으나, 대부분의 연구가 단순선형회귀모형에서 문제의 해결을 위한 제안에 불과할 뿐이다.

횡단면과 시계열 자료가 병합된 경시적 자료에 대한 패널회귀모형의 분석이 전통적인 횡단면 또는 시계열 자료 분석에 비해 시간에 따른 개체 특성의 변동과 개체에 따른 시간 특성의 변동 등 복합적인 문제에 대한 해답을 제공한다는 장점에도 불구하고 불완전 상태의 경시적 자료의 문제점을 극복하려는 연구는 매우 미진한 상태이다. 최근의 여러 패널모형연구(Balatgi, et al., 2001, 2002, 2003)들도 자료가 완전하다는 가정 아래서 다양한 추정방법과 검정방법들을 제안하고 이에 대한 성질을 연구하였을 뿐 자료가 불완전 상태에서의 추론의 문제는 다루지 못하였다. 따라서 본 연구에서는 관측된 경시적 자료가 불완전 상태(ill-posed)인 패널회귀모형에서 모수들에 대하여 GME방법을 이용하여 GME추정량을 유도하고 기존의 패널모형에서 이용하는 다양한 추정량들과의 효율성 비교를 통하여 GME 추정량의 강건한(robust) 성질에 대하여 다루고자한다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 GME 추정량을 유도하기 위해 GME 형식에 대해서 서술한다. 3장에서는 본 연구에서 다루고자 하는 패널회귀모형을 소개하고 전통적인 추정방법으로, 보통최소제곱(ordinary least sqaure, OLS)추정량, 일반화최소제곱(generalized least square, GLS) 추정량, 추정가능한 일반화최소제곱(Feasible GLS, FGLS) 추정량 등을 제시한다. 4장에서는 불완전 상태의 패널모형에서 GME 형식을 이용한 GME 추정량을 유도한다. 5장에서는 모의실험을 통하여 평균제곱오차(mean square error, MSE)를 이용하여 각 추정량의 효율성을 비교한다. 마지막으로 6장에서는 결론을 다룬다.

2. 최대엔트로피형식

엔트로피 측도(entropy measure)는 물리학에서 열역학적 상태를 정의하는 분포에서 정보를 측정하기 위해 Boltzman에 의해 처음으로 제안된 이후, 이것과 동일한 측도로 잡음(noise)을 포함하는 어떤 불확실성(uncertainty)을 측정하기 위해 Shannon(1948)에 의해 개발되고 제안되었다. 그리고 Jaynes(1957a, b)는 수학적으로 전통적인 과정으로는 다루어지지가 않는 역함수 문제의 추론을 위하여 Shannon이 제안한 엔트로피 거리(entropy metric)를 이용하여 최대엔트로피(maximum entropy, ME) 형식을 정의하였다. 최근에 이러한 잡음을 가진 역함수문제를 해결하기 위해 Jaynes의 ME형식을 확장하고 통계적 모형의 추정과 추론을 위하여 일반화최대엔트로피(generalized maximum entropy, GME) 형식

을 사용하고 있다. 특히 불완전 상태의 자료문제에 초점을 두고 Golan and Judge(1996), Golan, et al. (1996) 등이 활발히 연구 중에 있다.

2.1. 고전적 최대엔트로피형식

유한하고 이산형인 다음과 같은 선형모형을 고려해 보자

$$y = X\beta = Xp \tag{2.1}$$

여기서 y는 $T\times 1$인 종속변수 벡터이고, X는 $T\times K$인 $(T < K)$ 설명변수 행렬이며 비확률적이다. 그리고 빈도 $p = (p_1, ..., p_K)'$는 미지의 $K\times 1$인 모수 벡터이며 $\sum_{k=1}^{K} p_k = 1, p_k \geq 0$이다. 식 (2.1)에서 p를 추정하는 것이 목적이지만 관찰 자료의 수가 미지의 모수의 수보다 적기 때문에 p를 추정하기가 어렵다. 이에 대해 전통적으로 확률을 불확실성의 측도(measure of the uncertainty)로서 사용하기 때문에 Shannon(1948)은 엔트로피를 확률의 분포로서 정의하고 있다. 예를 들어 $x = (x_1, ..., x_K)'$를 확률변수라 하고, 각각의 확률을 $p = (p_1, ..., p_K)'$라 했을 때 p에 대한 분포의 엔트로피를 다음과 같이 정의한다.

$$H(p) \equiv -\sum_{k} p_k ln p_k, \quad 0 \cdot ln(0) = 0 \tag{2.2}$$

여기서 H는 사건들의 집합의 불확실성의 측도이며, p의 범위가 0과 1사이에서 증가하다 감소한다. 그리고 $\sum_{k=1}^{K} p_k = 1$이다. 식 (2.2)에서 엔트로피가 최대가 될 때의 p를 구해보면, $\hat{p}_1 = \hat{p}_2 = \cdots = \hat{p}_K = 1/K$일 때, 즉 각 확률들이 균일할 때, 엔트로피가 최대가 된다. 즉, 전통적으로 엔트로피의 개념은 알려지지 않은 확률의 분포를 선택할 때 사용되었다.

더불어 Jaynes(1957a, b)에 의해 일반화최대엔트로피(generalized maximum entropy, GME) 원칙은 자료의 정보가 확률 할당을 결정하는데 충분해야 한다. 이러한 접근을 위해 N번 시도에 K개 $(N_1, \cdots, N_K)$의 가능한 결과가 있는 다항분포를 고려해 보자.

$$f(N_1, \cdots, N_K) = \frac{N!}{N_1! \cdots N_K!} {p_1}^{N_1} \cdots {p_K}^{N_K} \tag{2.3}$$

여기서 $\sum_{k=1}^{K} N_k = N$, $\sum_{k=1}^{K} p_k = 1$이다. 식 (2.3)의 최대가능도추정량은 $N_k = Np_k$이며, N_k의 경우의 수는 다항계수 $W = \frac{N!}{Np_1! \cdots Np_K!}$로 표현 가능하다. 따라서 엔트로피가 다음과 같이 표현 가능하다(Golan, et al., 1996).

$$N^{-1} ln W \approx -\sum_{k=1}^{K} p_k ln p_k = H(p) \tag{2.4}$$

결론적으로 $H(p)$는 확률분포에서 불확실성의 측도이다.

확률 $p = (p_1, \cdots, p_K)'$의 분포에 대한 최대엔트로피(maximum entropy, ME) 형식을 정의하면 다음과 같다.

$$H(p) \equiv -\sum_{k=1}^{K} p_k ln p_k = -p' ln p, \quad 0 \cdot ln(0) = 0 \tag{2.5}$$

여기서 첫 번째 제약조건인 일치성(consistency)은 다음과 같다.

$$y = Xp \tag{2.6}$$

두 번째 제약조건인 정규화 가법성(normalization-additivity)은 다음과 같다.

$$p'1 = 1 \tag{2.7}$$

이러한 ME 형식은 Shannon(1948)과 Jaynes(1957a, b, 1984)이 제안을 했으며 Levine(1980), Levine and Tribus(1979)에 의해 발전되었다. 식 (2.5)를 최대로 하는 $\hat{p}$ 을 구하는 문제는 다음의 라그랑지안 함수(Lagrangian function)를 이용한다.

$$L = -p'lnp + \lambda'(y - Xp) + \mu(1 - p'1) \tag{2.8}$$

여기서 λ는 라그랑지안 상수(Lagrangian multipler)이다.

2.2. 일반화 최대엔트로피형식

본 절에서는 유한하고 이산형이며 오차항이 있는 다음과 같은 일반선형모형(general linear model, GLM)을 고려한다.

$$y = X\beta + e \tag{2.9}$$

여기서 y, X, β의 정의는 식 (2.1)과 동일하며, e는 $T \times 1$인 오차변수 벡터이다. 알려지지 않은 모수 $\beta = (\beta_1, \cdots, \beta_K)$와 오차 e가 제한된 사전정보를 가졌다고 가정한다. 예를 들어 모수 β와 오차 e가 임의의 범위(range)를 가지고 있다고 가정을 하면 모수 β와 오차 e가 사전 가중치나 유한 대응치(finite supports)를 가지는 이산확률변수로 구성되어 있다고 할 수 있다. 따라서 다음과 같이 모수 β와 오차 e를 재모수화(reparameterization) 할 수 있다. 각 β_k를 대응치 z와 $M(2 \leq M < \infty)$개의 가능한 결과치로 구성된 이산확률변수라고 가정한다. 만일 z_{k1}과 z_{kM}이 각각 모수 β_k 의 가능한 극한값들(상한값과 하한값)이라고 하면 β_k를 이러한 두점의 볼록 조합(convex combination)으로 표현할 수 있다. 즉, 다음과 같이 $M = 2$에 대해 각 극한값에 확률을 할당한다.

$$\beta_k = p_k z_{k1} + (1 - p_k) z_{kM}, \quad k = 1, \cdots, K \tag{2.10}$$

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

모수들을 일반적으로 표현하기 위하여 z_k를 $\pounds$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 $\beta \in int(\pounds)$라도 다음과 같이 행렬로 표현할 수 있다.

$$\beta = Zp = \begin{bmatrix} z_1 & 0 & \cdots & 0 \\ 0 & z_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & z_K \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_K \end{bmatrix} \tag{2.11}$$

여기서 Z는 $(K \times K)$행렬이고, p는 K차원의 가중치벡터이다.

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

$$e_t = w_t v_{t1} + (1 - w_t) v_{tJ} \tag{2.12}$$

여기서 $w_t \in [0,1]$이다. 오차항 또한 다음과 같이 행렬의 형태로 재모수화할 수 있다.

$$e = Vw = \begin{bmatrix} v_1 & 0 & \cdots & 0 \\ 0 & v_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & v_T \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_T \end{bmatrix} \tag{2.13}$$

여기서 V는 $(T \times T)$행렬이고, w는 T차원의 가중치벡터이다. 식 (2.11)과 식(2.13)에 의해 재모수화된 $\beta = Zp$와 $e = Vw$를 사용함으로써 Judge and Golan(1992)은 GLM을 다음과 같이 표현하였다.

$$y = X\beta + e = XZp + Vw \tag{2.14}$$

그리고, 식 (2.14)을 포괄적 선형모형(generic linear model)으로 다음과 같이 재정의할 수 있다.

$$\alpha = \Gamma\beta + \epsilon \tag{2.15}$$

따라서 식 (2.15)를 GLM의 포괄적 GME 형식으로 다음과 같이 재표현할 수 있다.

$$\alpha = \Gamma Zp + Vw \tag{2.16}$$

다음으로 일반화 최대엔트로피(GME) 형식은 다음과 같다.

$$H(p,w) = -p'ln(p) - w'ln(w) \tag{2.17}$$

여기서 제약조건은 일치성조건으로 $\alpha = \Gamma Zp + Vw$ 이며, 정규화 가법성조건으로 $i_K = (I_K \otimes i'_M)p$ 와 $i_T = (I_T \otimes i'_J)w$ 이며 $\otimes$는 크로네커곱(Kronecker product)을 나타낸다. 식 (2.17)에서 최대값 $\hat{p}, \hat{w}$를 구하기 위해 라그랑지안 함수를 다음과 같이 정의한다.

$$\begin{aligned} L \;=\; & -p'lnp - w'lnw + \lambda'[\alpha - \Gamma Zp - Vw] + \\ & \theta'[i_K - (I_K \otimes i'_M)p] + \tau'[i_T - (I_T \otimes i'_J)w] \end{aligned} \tag{2.18}$$

여기서 I_K는 K차원의 단위행렬이고, i_T는 모든 원소가 1인 크기가 T인 벡터이다. $p, w, \lambda, \theta, \tau$ 각각의 값에 대한 1차 미분을 하여 $\tilde{p}$ 추정량을 구하면 다음과 같다.

$$\tilde{p} = exp(-Z'\Gamma'\lambda) \odot \{(I_K \otimes i_M i'_M) exp(-Z'\Gamma'\lambda)\}^{-1} \tag{2.19}$$

여기서 ⊙는 하다마드곱(Hadamard product)이다. 이를 통해 모수를 추정하면 다음과 같다.

$$\begin{aligned}\tilde{\beta}_{GME} &= Z\tilde{p} \\ &= Z \cdot exp(-Z'\Gamma'\lambda) \odot \{(I_K \otimes i_M i_M')exp(-Z'\Gamma'\lambda)\}^{-1} \end{aligned} \tag{2.20}$$

식 (2.20)에서 k번째 $\tilde{\beta}_{GME}$를 구해보면 다음과 같다.

$$\tilde{\beta}_{GME(k)} = \sum_{m=1}^{M} z_{km}\tilde{p}_{km} = \sum_{m=1}^{M} z_{km}\Big[\frac{exp\big(-z_{km}\sum_{t=1}^{T}\gamma_{kt}\lambda\big)}{\Omega_k(\tilde{\lambda})}\Big] \tag{2.21}$$

여기서 $\Omega_k(\tilde{\lambda}) = \sum_{m=1}^{M} exp\big(-z_{km}(\sum_{t=1}^{T}\gamma_{kt}\lambda)\big)$.

3. 패널회귀모형

횡단면 자료와 시계열자료가 병합된 다음과 같은 패널회귀모형을 고려한다.

$$y_{it} = x_{it}'\beta + u_{it}, \quad i = 1, \cdots, N \quad t = 1, \cdots, T \tag{3.1}$$

여기서 y_{it}는 i번째 개체(개인, 가구, 국가 등)의 시점 t에서의 관측치를 나타내는 반응값이고, x_{it}는 K개의 변수로 이루어진 설명변수 벡터이다. 모형 (3.1)에서 오차항 u_{it}는 다음과 같은 일원오차성분(one way error components)을 갖는다고 가정하자.

$$u_{it} = \mu_i + e_{it}, \quad i = 1, \cdots, N, \quad t = 1, \cdots, T \tag{3.2}$$

여기서 μ_i가 관측될 수 없는 개체효과(individual effect)를 나타내는 확률변수이며, e_{it}는 나머지 오차항을 나타낸다. μ_i, e_{it}는 서로 독립이며 각각 $\mu_i \sim i.i.d.(0, {\sigma_\mu}^2), e_{it} \sim i.i.d.(0, {\sigma_e}^2)$이라고 가정한다. 모형 (3.1)을 행렬을 이용하여 나타내면 다음과 같다.

$$y = X\beta + u \tag{3.3}$$

y는 $(NT \times 1)$인 종속변수 벡터, X는 $(NT \times K)$인 독립변수 행렬이며, β는 $K \times 1$인 회귀계수 벡터이다. 또한 식 (3.2)의 오차항을 행렬로 표현하면 다음과 같다.

$$u = (I_N \otimes i_T)\mu + e \tag{3.4}$$

여기서 $\mu = (\mu_1, \cdots, \mu_N)', e = (e_{11}, \cdots, e_{NT})'$이다. 식 (3.4)에 대한 오차항의 분산-공분산 행렬을 다음과 같이 구해진다.

$$\begin{aligned}\Omega &= E(uu') = (I_N \otimes i_T)E(\mu\mu')(I_N \otimes i_T)' + E(ee') \\ &= {\sigma_\mu}^2(I_N \otimes J_T) + {\sigma_e}^2(I_N \otimes I_T) \end{aligned} \tag{3.5}$$

여기서 J_T는 모든 원소가 1인 $(T \times T)$행렬이다. 식 (3.5)에서 $\bar{J}_T = J_T/T$라 하고 $E_T = I_T - \bar{J}_T$라 할 때, I_T를 $E_T + \bar{J}_T$로 치환하면 식 (3.5)의 Ω는 다음과 같이 표현할 수 있다.

$$\Omega = {\sigma_1}^2 P + {\sigma_e}^2 Q \tag{3.6}$$

여기서 $\sigma_1^2 = T\sigma_\mu^2 + \sigma_e^2$이고, $P = (I_N \otimes \bar{J}_T)$, $Q = (I_N \otimes E_T)$이다. Wansbeek and Kapteyn(1982)의 결과를 이용하면 Ω^r은 다음과 같이 표현된다.

$$\Omega^r = (\sigma_1^2)^r P + (\sigma_e^2)^r Q \tag{3.7}$$

r는 임의의 상수이고, 식 (3.7)을 이용하면 Ω의 역행렬과 변환행렬 $\Omega^{-1/2}$를 다음과 같이 구할 수 있다.

$$\begin{aligned} \Omega^{-1} &= (\sigma_1^2)^{-1} P + (\sigma_e^2)^{-1} Q \\ \Omega^{-1/2} &= \frac{1}{\sigma_1} P + \frac{1}{\sigma_e} Q \end{aligned} \tag{3.8}$$

3.1. OLS, GLS 추정량

모형 (3.3)에서 β에 대한 보통최소제곱추정량(OLS추정량)은 다음과 같이 구해진다.

$$\widehat{\beta}_{OLS} = (X'X)^{-1} X'y \tag{3.9}$$

OLS추정량은 불편추정량이며 일치추정량이지만 분산성분들의 존재를 무시하였기 때문에 OLS추정량의 효율은 떨어질 것이다. 또한 OLS 추정에 의하여 유도된 표준오차는 편향된다는 사실이 잘 알려져 있다(Moulton, 1986). OLS추정량에 의한 잔차는 $\widehat{u}_{OLS} = y - X\widehat{\beta}_{OLS}$이다. 만일 식 (3.5)의 분산성분들이 알려져있다면, β의 GLS추정량은 다음과 같다.

$$\widehat{\beta}_{GLS} = (X'\Omega^{-1}X)^{-1} X'\Omega^{-1} y \tag{3.10}$$

그러나, 현실적으로 분산성분들은 알려지지 않은 경우가 대부분이므로 GLS추정량은 단지 이론적인 추정량에 불과하나 다른 추정량과의 비교시 사용될 수 있다.

3.2. FGLS 추정량

만일 식 (3.5)에서 분산성분들이 알려져 있지 않으면, 분산성분들을 먼저 추정하고 이들을 식 (3.10)에 대입하여 회귀계수를 추정하는 방법으로 얻어지는 추정량을 FGLS(feasible GLS)추정량이라 한다.

$$\widehat{\beta}_{FGLS} = (X'\widehat{\Omega}^{-1}X)^{-1} X'\widehat{\Omega}^{-1} y \tag{3.11}$$

여기서 $\widehat{\Omega} = \widehat{\sigma}_1^2 P + \widehat{\sigma}_e^2 Q$이다. FGLS추정량은 점근적으로 GLS추정량과 동일한 성질을 갖게 되지만 일반적으로 소표본인 경우에는 정확한 성질이 알려져 있지 않다(Judge et al., 1988). 분산성분들을 추정하는 방법으로 Amemiya(1971)와 Swamy and Arora(1972)에 의해 제안된 방법으로 분산성분을 추정하여 회귀계수를 추정하고자 한다.

Amemiya(1971)는 분산성분들을 추정하기 위하여 개체효과를 제거시킨 LSDV(least square dummy variables) 잔차를 이용하였다. 이 경우 얻어지는 회귀계수에 대한 추정량(AM추정량)은 다음과 같다.

$$\begin{aligned} \widetilde{\beta}_{AM} &= (X'\widehat{\Omega}^{-1}X)^{-1} X'\widehat{\Omega}^{-1} y \\ &= \left[(X' \Big(\frac{1}{\widehat{\sigma}_1^2} P + \frac{1}{\widehat{\sigma}_e^2} Q \Big)^{-1} X \right]^{-1} X' \Big(\frac{1}{\widehat{\sigma}_1^2} P + \frac{1}{\widehat{\sigma}_e^2} Q \Big)^{-1} y \end{aligned} \tag{3.12}$$

여기서 $\widehat{\sigma}_1^2 = \dfrac{u'Pu}{tr(P)} = T\sum_{i=1}^{N}\widehat{u}_{i.}^2/N$, $\widehat{\sigma}_e^2 = \dfrac{u'Qu}{tr(Q)} = \sum_{i=1}^{N}\sum_{t=1}^{T}(u_{it}-\widehat{u}_{i.})^2/N(T-1), \widehat{u} = \widehat{u}_{LSDV}$.

Swamy and Arora(1972)는 분산성분을 추정하기 위하여 내부변이와 개체변이를 이용하는 두 개의 회귀모형을 적합하였으며, 각 모형의 평균제곱 오차를 이용하여 분산성분을 추정하였다. 이 경우 얻어지는 회귀계수에 대한 추정량(SA추정량)은 다음과 같다.

$$\begin{aligned}\widetilde{\beta}_{SW} &= (X'\widehat{\widehat{\Omega}}^{-1}X)^{-1}X'\widehat{\widehat{\Omega}}^{-1}y \\ &= \Big[(X'\Big(\frac{1}{\widehat{\widehat{\sigma}}_1^2}P+\frac{1}{\widehat{\widehat{\sigma}}_e^2}Q\Big)^{-1}X\Big]^{-1}X'\Big(\frac{1}{\widehat{\widehat{\sigma}}_1^2}P+\frac{1}{\widehat{\widehat{\sigma}}_e^2}Q\Big)^{-1}y \qquad (3.13)\end{aligned}$$

여기서 $\widehat{\widehat{\sigma}}_1^2 = \dfrac{y'Py - y'PX(X'PX)^{-1}X'Py}{N-K-1}$, $\widehat{\widehat{\sigma}}_e^2 = \dfrac{y'Qy - y'QX(X'QX)^{-1}X'Qy}{N(T-1)-K}$.

4. 회귀계수에 대한 GME 추정량

GME 추정량을 구하기 위해 회귀계수와 개체효과, 그리고 오차항을 재모수화한다. 첫 번째로 회귀계수를 재모수화하면 다음과 같다.

$$\beta = Zp = \begin{bmatrix} z_1' & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & z_k' & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \cdots & z_K' \end{bmatrix}\begin{bmatrix} p_1 \\ \vdots \\ p_k \\ \vdots \\ p_K \end{bmatrix} \qquad (4.1)$$

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

다음으로 개체효과를 재모수화하면 다음과 같다.

$$\mu = Fg = \begin{bmatrix} f_1' & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & f_n' & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \cdots & f_N' \end{bmatrix}\begin{bmatrix} g_1 \\ \vdots \\ g_n \\ \vdots \\ g_N \end{bmatrix} \qquad (4.2)$$

여기서 $\mu_n = \sum_{i=1}^{I} f_{ni}g_{ni}, n=1,\cdots,N, f_n = (f_{n1}, f_{n2}, \cdots, f_{nI})', g_n = (g_{n1}, g_{n2}, \cdots, g_{nI})'$.

마지막으로 나머지 오차항을 재모수화하면 다음과 같다.

$$e = Vw = \begin{bmatrix} v'_{11} & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & v'_{nt} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \cdots & v'_{NT} \end{bmatrix} \begin{bmatrix} w_{11} \\ \vdots \\ w_{nt} \\ \vdots \\ w_{NT} \end{bmatrix} \tag{4.3}$$

여기서 $e_{nt} = \sum_{j=1}^{J} v_{ntj} w_{ntj}$, $v_{nt} = (v_{nt1}, v_{nt2}, \cdots, v_{ntJ})$, $w_{nt} = (w_{nt1}, w_{nt2}, \cdots, w_{ntJ})$.

다음으로 GME형식을 이용하여 모수에 대한 GME 추정량을 구한다.

$$max_{p,g,w} H(p,g,w) = -p'ln(p) - g'ln(g) - w'ln(w) \tag{4.4}$$

첫 번째 조건은 일치성으로 다음과 같다.

$$y = XZp + (I_N \otimes i_T)Fg + Vw \tag{4.5}$$

두 번째 조건은 정규화 가법성으로 다음과 같다.

$$i_K = (I_K \otimes i'_M)p, \quad i_{NT} = (I_{NT} \otimes i'_J)w, \quad i_N = (I_N \otimes i'_I)g \tag{4.6}$$

이제 GME 추정량 β_{GME}를 유도해 보자. 먼저 라그랑지안 방정식을 다음과 같이 정의한다.

$$\begin{aligned} L \;=\; & -p'lnp - w'lnw - g'lng + \lambda'[y - XZp - (I_N \otimes i_T)Fg - Vw] + \\ & \theta'[i_K - (I_K \otimes i'_M)p] + \tau'[i_{NT} - (I_{NT} \otimes i'_J)w] + \gamma'[i_N - (I_N \otimes i'_I)g] \end{aligned} \tag{4.7}$$

식 (4.7)에서 $p, w, g, \lambda, \theta, \tau, \gamma$ 각각의 값에 대해 1차 미분하면 다음과 같다.

$$\nabla_p L = -lnp - i_{KM} - Z'X'\lambda - (I_K \otimes i_M)\theta = 0 \tag{4.8}$$

$$\nabla_w L = -lnw - i_{NTJ} - V'\lambda - (I_{NT} \otimes i_J)\tau = 0 \tag{4.9}$$

$$\nabla_g L = -lng - i_{NI} - F'(I_N \otimes i'_T)\lambda - (I_N \otimes i_I)\gamma = 0 \tag{4.10}$$

$$\nabla_\lambda L = y - XZp - (I_N \otimes i_T)Fg - Vw = 0 \tag{4.11}$$

$$\nabla_\theta L = i_K - (I_K \otimes i'_M)p = 0 \tag{4.12}$$

$$\nabla_\tau L = i_{NT} - (I_{NT} \otimes i'_J)w = 0 \tag{4.13}$$

$$\nabla_\gamma L = \iota_N - (I_N \otimes i'_I)g = 0 \tag{4.14}$$

여기서 $\tilde{p}$, $\tilde{w}$, $\tilde{g}$ 추정량을 구해보면 다음과 같다. 식 (4.8)과 (4.12)에 의해

$$\tilde{p} = exp(-Z'X'\lambda) \odot \{(I_K \otimes i_M i'_M)exp(-Z'X'\lambda)\}^{-1} \tag{4.15}$$

이고, 식 (4.9)와 (4.13)에 의해

$$\tilde{w} = exp(-V'\lambda) \odot \{(I_{NT} \otimes i_J i'_J)exp(-V'\lambda)\}^{-1} \tag{4.16}$$

이고, 식 (4.10)과 (4.14)에 의해

$$\tilde{g} = exp(-F'(I_N \otimes i_T')\lambda) \odot \{(I_N \otimes i_I i_I')exp(-F'(I_N \otimes i_T')\lambda)\}^{-1}. \tag{4.17}$$

마지막으로 회귀계수를 추정하면 다음과 같다.

$$\tilde{\beta} = Z\tilde{p} = Z \cdot exp(-Z'X'\lambda) \odot \{(I_K \otimes i_M i_M')exp(-Z'X'\lambda)\}^{-1} \tag{4.18}$$

식 (4.18)에서 k번째 β_{GME}를 구해면 다음과 같다.

$$\tilde{\beta}_{GME(k)} = \sum_{m=1}^{M} z_{km}\tilde{p}_{km} = \sum_{m=1}^{M} z_{km}\Big[\frac{exp\big(-z_{km}\sum_{n=1}^{N}\sum_{t=1}^{T}x_{knt}\lambda\big)}{\Omega_k(\tilde{\lambda})}\Big] \tag{4.19}$$

여기서 $\Omega_k(\tilde{\lambda}) = \sum_{m=1}^{M} exp\Big[-z_{km}\Big(\sum_{n=1}^{N}\sum_{t=1}^{T}x_{knt}\lambda\Big)\Big]$.

5. 모의실험

5.1. 모의실험 방법

앞 장에서 유도한 회귀계수에 대한 여러 추정량들의 효율성을 비교하기 위하여 모의실험을 실시하였다. 모의실험에 사용된 모형은 다음과 같은 패널회귀모형이다.

$$\begin{aligned} y_{it} &= \beta x_{it} + u_{it}, \quad i = 1, 2, \cdots, N, \quad t = 1, 2, \cdots, T, \\ u_{it} &= \mu_i + e_{it} \end{aligned} \tag{5.1}$$

여기서 y는 (125×1)인 종속변수 벡터이며, $N = 25$이고 $T = 5$이다. X는 (125×4)인 독립변수 벡터이며, β는 (4×1)인 회귀계수 벡터이고, μ는 (125×1), e는 (125×1)인 오차벡터로써 정규분포 $N(0,1)$에서 발생시켰다. 식 (5.1)에서 원하는 조건수 $c(X'X) = \mu$ 를 가지는 X의 계획행렬을 구하기 위해 X의 SVD(singular value decomposition), $X = QLR$을 구한다. 즉, 조건수 $c(X_a'X_a) = \mu$ 에 따라 X가 $X_a = QL_aR$로 새로이 구해진다. 이때, L에서의 고유값은 4차원의 벡터인

$$a = \Big[\sqrt{\frac{2}{1+\mu}}, 1, 1, \sqrt{\frac{2\mu}{1+\mu}}\Big] \tag{5.2}$$

로 대치된다(Belsley, 1991). 또한 $\beta = [2, 1, -1, 2]$에 대해서 관찰치 y를 형성하기 위해 $X_a\beta$와 25개의 $N(0,1)$ 분포를 가지는 μ와 125개의 $N(0,1)$ 분포를 가지는 오차를 구한다. GME 추정량의 모수 대응치는 $z_k = [-10, -5, 0, 5, 10]$, 개체 대응치는 $f_n = [-3, 0, 3]$, 오차 대응치는 $v_{it} = [-3, 0, 3]$로 정한다. 전체분산은 $\sigma_\mu^2 + \sigma_e^2 = 20$으로 고정하여 σ_μ^2를 0에서 16까지 변화시켰다 (σ_e^2은 20에서 4까지). 그리고, 조건수 $c(X'X) = \mu$의 값은 $N(0,1)$으로 부터 구한 자료인 '1' 부터, 클수록 자료의 왜곡상태가 심한 '10', '50', '100' 까지 총 4 가지 경우로 정하였다. 각 경우의 반복횟수는 1000회로 하였다.

5.2. 모의실험의 결과

표 5.1은 σ_μ^2과 σ_e^2의 값에 따라, 그리고 조건수 $\mu(=1,10,50,100)$의 값에 따라 GME 추정량과 다른 추정량과의 MSE를 비교한 결과이다.

표 5.1: 추정량의 평균제곱오차

σ_μ^2	σ_e^2	Condition	OLS-E	GLS-E	AM-E	SA-E	GME-E
4	16	1	0.166	0.149	0.150	0.151	2.500
		10	41.470	37.394	38.081	38.322	2.501
		50	140.330	128.720	131.573	131.623	2.502
		100	270.630	238.217	238.286	241.437	2.501
8	16	1	0.157	0.114	0.114	0.114	2.500
		10	39.369	28.166	28.282	28.329	2.502
		50	152.310	104.653	104.989	105.383	2.503
		100	271.258	186.737	188.687	188.711	2.500
12	8	1	0.163	0.081	0.081	0.081	2.500
		10	40.852	19.385	19.412	19.451	2.508
		50	140.655	68.929	68.974	68.940	2.505
		100	268.274	131.829	132.270	132.459	2.503
16	4	1	0.172	0.043	0.043	0.043	2.500
		10	40.243	10.233	10.250	10.236	2.502
		50	143.381	37.175	37.241	37.218	2.502
		100	261.060	63.670	63.790	63.773	2.502

표 5.1을 보면 σ_μ^2과 σ_e^2의 값에 관계없이 모든 경우에 조건수 μ의 값이 1일 때 GME 추정량이 상당히 효율성이 떨어지는 경향을 보여주고 있다. 예를 들어 $\mu=1$, $\sigma_\mu^2=4$, $\sigma_e^2=16$ 일 때 OLS 추정치는 0.170, GLS 추정치는 0.152, 그리고, AM 추정치는 0.152, SW 추정치는 0.154인데 비해 GME 추정치는 1.951로 상당히 효율성이 떨어짐을 알 수 있다. 이러한 현상은 GME 추정량은 자료의 불완전 상태를 고려한 추정량이므로 완전한 자료일 경우 GME 추정량이 별로 좋은 추정량이 될 수 없음을 알 수 있다. 그러나, 자료의 왜곡상태가 심할 경우에는 GME 추정량이 상당히 효율성이 좋음을 알 수 있다. 예를 들어, $\mu=50$, $\sigma_\mu^2=4$, $\sigma_e^2=16$ 일 때 OLS 추정치는 141.970, GLS 추정치는 126.117, 그리고, AM 추정치는 126.793, SW 추정치는 127.039 로 상당히 큰 MSE값을 주는데 비해 GME 추정치는 2.513으로 상당히 작은 값으로 효율성이 좋음을 알 수 있다. 자료의 상태가 불완전 하더라도 GLS추정량과 FGLS 추정량들(AM, SA)과의 MSE값이 별차이가 없이 나타나는 점에 유의해야 한다. 즉 자료의 왜곡상태를 고려하지 않고 FGLS를 이용하여 회귀계수를 추정한다면 통계적 추론 또한 잘못된 결론이 도출될 수 있다. 또한, 모의실험을 할 때마다 OLS, GLS, FGLS 추정치 모두가 상당히 큰 폭으로 다른 값을 주었으나, GME 추정치는 거의 변동 없이 상당히 안정적인 MSE값(2.5 근처)을 주었다. 즉, GME 추정치는 엔트로피 값이 최대일 때의 추정치이기 때

문에 모든 경우의 추정치가 거의 균일하다. 결론적으로 패널회귀모형에 있어서 자료의 왜곡상태가 심할 경우, 즉 자료가 매우 불완전 상태인 경우 GME 형식을 이용한 추정량은 기존의 여러 다른 추정량에 비하여 상당히 효율적인 추정량이 될 수 있음을 알 수 있으며, 특히 불완전상태가 심할수록 robust한 추정량임을 보이고 있다.

6. 결론

패널회귀모형에 있어서 자료가 완전한 상태인지 여부는 통계적 추론을 위하여 매우 중요하다. 특히, 비실험자료가 주어졌을 때, 역함수의 해가 정의되지 않는 경우, 또는 추정치의 분산이 크고 정도가 낮을 때 이와 같은 불완전 상태의 문제가 발생하는데 만약 이러한 문제를 간과하고 자료를 분석하게 되면 올바른 통계적인 추론을 할 수 없게 된다. 그러므로 패널회귀모형에서 자료가 불완전 상태일 때 적절한 통계적 추정을 위하여 본 연구에서 제안하고 있는 GME 추정방법을 고려해 보는 것이 필요하다. 모의실험을 통하여 각 추정치의 값을 구하는 과정에서 GME 추정량은 반복적인 수행을 통해 라그랑지안 배수 추정치를 구해야 하기 때문에 다소 복잡한 과정이 필요하였다. 그러나 모의실험 결과에서 알 수 있듯이 GME 추정량은 OLS, GLS, FGLS 추정량 등 다른 추정량들과 효율성을 비교를 통하여 상당히 안정적이고 매우 효율적인 추정량임을 보여주었다. 따라서 패널회귀모형에서 자료의 형태가 불완전 상태인 경우에는 GME 추정량을 사용하는 것이 적절할 것으로 생각된다.

참고문헌

Amemiya, T. (1971). The estimation of the variances in a variance components model, *International Econometric Review*, **12**, 1-13.

Baltagi, B. H., Song, S. H. and Jung, B. C. (2001). The unbalanced nested error component regression model, *Journal of Econometrics*, **101**, 357-381.

Baltagi, B. H., Song, S. H. and Jung, B. C. (2002). Simple LM tests for unbalanced nested error component regression model, *Econometric Reviews*, **21**, 167-187.

Baltagi, B. H., Song, S. H. and Koh, W. (2003). Testing panel data regression model with spatial error correlation, *Journal of Econometrics*, **117**, 123-150.

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

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

Golan, A., Judge, G. and Robinson, S. (1994). Recovering information in the case of partial multisectorial economic data, *Review of Economics and Statistics*, **76**, 541-549.

Golan, A. and Judge, G. (1996). Recovering information in the case of underdetermined problems and incomplete data, *Journal of Statistical Planning and Inference*, **49**, 127-136.

Golan, A., Judge, G., and Miller, D. (1996). *Maximum Entropy Econometrics: Robust Estimation with Limited Data*, John Wiley, New York.

Jaynes, E. T. (1957a). Information theory and statistical mechanics, *Physics Review*, **106**, 620-30.

Jaynes, E. T. (1957b). Information theory and statistical mechanics II, *Physics Review*, **108**, 171-90.

Jaynes, E. T. (1984). Prior information and ambiguity in inverse problems, In D. W. McLaughlin (Ed.) *Inverse problems*, p.151-66, SIAM Proceedings, American Mathematical Society, Providence, RI.

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.

Judge, G. G., Hill, R. C., Griffiths, W. E., Lutkepohl, H. and Lee, T. C. (1988). *Introduction to the Theory and Practice of Econometrics*, John Wiley, New York.

Levine, R. D. (1980). An information theoretical approach to inversion problems, *Journal of Physics*, **13**, 91-108.

Levine, R. D. and Tribus, M. (1979). *The Maximum Entropy Formalism*, MIT Press, Cambridge.

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

Shannon, C. E. (1948). A mathematical theory of communication, *Bell system Technical Journal*, **27**, 379-423.

Swamy, P. A. V. B. and Arora, S. S. (1972). The exact finite sample properties of the estimators of coefficients in the error components regression models, *Econometrica*, **40**, 261-275.

Wansbeek, T. J. and Kapteyn, A. (1982). A simple way to obtain the spectral decomposition of variance components model for balanced data, *Communication in Statistics*, **11**, 2105-2112.

[2006년 2월 접수, 2006년 8월 채택]

A Study of Generalized Maximum Entropy Estimator for the Panel Regression Model*

Seuck Heun Song[1] Soo Young Cheon[2]

ABSTRACT

This paper considers a panel regression model with ill-posed data and proposes the generalized maximum entropy(GME) estimator of the unknown parameters. These are natural extensions from the biometrics, statistics and econometrics literature. The performance of this estimator is investigated by using of Monte Carlo experiments. The results indicate that the GME method performs the best in estimating the unknown parameters.

Keywords: Panel Regression Model, Information Recovery, ME Estimation, GME Estimation

* This work was supported by Korea Research Foundation Grant(KRF-2004-C00073-101102)

1) Professor, Dept of Statistics, Korea University, Annam-Dong 5-1, Seoul 136-701, Korea
E-mail: ssong@korea.ac.kr

2) Graduate student, Dept of Statistics, Korea University, Annam-Dong 5-1, Seoul 136-701, Korea
E-mail: sycheon@korea.ac.kr