한국통계학회논문집
2012, 19권, 3호, 345–358

DOI: http://dx.doi.org/10.5351/CKSS.2012.19.3.345

비정규분포를 이용한 표본선택 모형 추정: 자동차 보유와 유지비용에 관한 실증분석

최필선[a], 민인식[1,b]

[a]건국대학교 국제무역학과, [b]경희대학교 경제학부

요 약

표본선택 모형을 최우추정법으로 추정할 때 오차항의 분포를 제대로 가정하는 것이 매우 중요하다. 표본선택 모형의 선택 방정식과 본 방정식의 오차항 분포를 일반적으로 이변량 정규분포로 가정하지만, 이 가정이 오차항의 실제 분포를 과도하게 제약할 가능성이 있다. 본 연구는 표본선택 모형의 오차항 분포로 S_U-정규분포를 도입한다. S_U-정규분포는 분포의 비대칭성과 초과첨도를 허용한다는 측면에서 정규분포보다 훨씬 유연하면서, 동시에 정규분포를 극한분포의 형태로 포함하고 있다. 또한 정규분포처럼 다변량 분포함수가 존재하기 때문에 표본선택 모형과 같은 다변량 모형에서도 활용할 수 있다. 본 논문은 S_U-정규분포를 이용한 표본선택 모형에서 로그우도 함수와 조건부 기댓값을 도출하고, 시뮬레이션을 통해 정규분포 모형과 추정성과를 비교한다. 또한 자동차 보유 가구들의 자동차 유지비에 관한 실제 데이터를 이용하여 S_U-정규분포 표본선택 모형의 추정결과를 제시한다.

주요용어: 표본선택 모형, 최우추정, 선택편의, S_U-정규분포.

1. 서론

로그우도함수를 최대화하여 모수를 추정하는 최우추정법(maximum likelihood estimation)은 사회과학의 다양한 회귀분석에서 이용되고 있다. 특히 로짓(Logit)이나 토빗(Tobit) 모형처럼 제한된 종속변수(limited dependent variable)를 사용하는 모형에서는 최우추정법이 주로 사용된다. 이들 모형에서 최우추정량이 일치추정량(consistent estimator)이 되기 위해서는 오차항의 분포가 올바르게 가정되어야 한다.

본 논문에서는 Heckman (1979)에 의해 소개된 표본선택(sample selection) 모형의 최우추정량에 대해서 논의한다. 표본선택 모형은 1단계 의사결정인 선택 방정식(selection equation)과 1단계에서 선택된 그룹의 2단계 의사결정인 본 방정식(main equation)으로 구성되어 있다. 그런데 2단계에서 관찰된 표본이 무작위 표본이 아니라 1단계 의사결정과 연관되어 있다면 선택편의(selection bias)를 고려하여 모수를 추정해야 한다는 것이 Heckman의 주장이다. 이러한 선택편의를 무시하고 각 방정식을 따로 추정하면 본 방정식에 있는 모수 추정량이 일치추정량이 되지 못한다는 것이다.

표본선택 모형은 경제학의 다양한 이슈에 적용되어 왔다. 취업여부의 의사결정과 취업한 사람들의 임금결정 모형, 주택의 소유여부와 소유한 주택의 가격결정 모형 등이 그 예이다. 표본선택 모형은 선택 방정식과 본 방정식의 상관성을 가정하고, 두 방정식의 모수에 대하여 최우추정량을 구한다. 이 경우 일반적으로 두 방정식의 오차항 분포를 이변량 정규분포로 가정한다. 하지만 Manski (1989)에 따르

이 논문은 2012년도 경희대학교 연구년 지원에 의한 결과임.
[1] 교신저자: (130-701) 서울시 동대문구 회기동 1번지, 경희대학교 정경대학 경제학과, 부교수.
E-mail: imin@khu.ac.kr

면 오차항 분포 가정에 따라 추정치가 많이 달라지는데, 정규분포 가정이 오차항의 실제 분포를 과도하게 제약할 가능성이 있다.

본 연구에서는 표본선택 모형의 오차항 분포로 S_U-정규분포를 도입한다. S_U-정규분포는 분포의 비대칭성(asymmetry)과 초과첨도(excess kurtosis)를 허용한다는 측면에서 정규분포보다 훨씬 유연하면서, 동시에 정규분포를 극한분포(limiting distribution)의 형태로 포함하고 있다. 또한 S_U-정규 확률변수는 정규 확률변수의 특정 변환으로 정의되기 때문에 정규분포처럼 다변량(multivariate) 분포함수가 존재한다. 따라서 표본선택 모형처럼 다변량모형에서도 활용할 수 있다는 장점이 있다. 본 논문은 표본선택 모형에서 정규분포의 제약성을 완화할 수 있는 유연한 분포함수를 가정한 후 로그우도함수와 조건부 기댓값을 도출하고, 시뮬레이션을 통해 정규분포 모형과 추정성과를 비교한다. 또한 실제 데이터를 이용하여 S_U-정규분포를 활용한 표본선택 모형의 추정결과를 제시한다. 실제 데이터 분석은 1단계 선택 방정식에서는 가구의 자동차 보유 여부를 종속변수로 두었다. 또한 2단계 본 방정식에서는 자동차 보유 가구들의 자동차 유지비용을 종속변수로 두고 표본선택 모형을 설정하였다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 Heckman이 소개한 정규분포 모형을 간단히 정리하고, 저자들이 도출한 S_U-정규분포 모형의 로그우도함수를 제시한다. 3장에서는 시뮬레이션을 통해 두 가지 분포 가정 하에서 구한 최우추정치의 성과를 비교한다. 4장에서는 가구별 자동차 보유 여부와 자동차 유지비 데이터를 이용하여 표본선택 모형을 추정하고 그 결과를 해석한다. 5장에서는 연구의 결과를 요약하고 결론을 제시한다.

2. 로그우도함수와 조건부 기댓값

본 절에서는 모수적(parametric) 분포함수를 가정한 표본선택 모형의 추정을 위한 로그우도함수에 대해 논의한다. 이변량 정규분포 모형에서의 최우추정량을 소개하고, 본 연구에서 새롭게 제시하는 이변량 S_U-정규분포 모형에서의 최우추정량을 도출한다.

2.1. 이변량 정규분포 모형

표본선택 모형에서 1단계의 선택 방정식은 전체 모집단이 무작위 표본으로 구성되어 있고, 종속변수는 이항(binary) 변수 타입이며, 0과 1의 의사결정은 설명변수 x_1에 의존한다고 가정된다.

$$y_{1i}^* = x_{1i}\beta + u_{1i}, \quad i = 1, \ldots, n, \tag{2.1}$$

여기에서 y_1^*는 관찰되지 않으며, 그 대신 관찰되는 종속변수 y_1은 다음과 같이 가정된다.

$$y_{1i} = 1, \quad y_{1i}^* > 0\text{일 때},$$
$$y_{1i} = 0, \quad y_{1i}^* \le 0\text{일 때}.$$

제 2단계의 본 방정식에서 종속변수 y_2는 연속형 변수이다. 그런데 2단계에서 관찰되는 것은 전체 모집단의 무작위 표본이 아니라 $y_{1i} = 1$인 표본에 대해서만 y_{2i}의 값이 관찰되며, $y_{1i} = 0$인 경우에는 관찰되지 않는다. 본 방정식은 다음과 같이 쓸 수 있다.

$$y_{2i} = x_{2i}\gamma + u_{2i}. \tag{2.2}$$

식 (2.1)과 (2.2)의 오차항이 서로 독립이 아니고 상관관계를 가지면, 식 (2.2)의 OLS 추정량은 일치추

정량이 되지 못한다. 식 (2.2)의 조건부 기댓값은 다음과 같이 쓸 수 있다.

$$\begin{aligned} E\left(y_{2i}|x_{2i}, y_{1i}^* > 0\right) &= E(x_{2i}\gamma + u_{2i}|x_{1i}\beta + u_{1i} > 0) \\ &= x_{2i}\gamma + E(u_{2i}|u_{1i} > -x_{1i}\beta). \end{aligned} \tag{2.3}$$

식 (2.3)에서 u_{2i}와 u_{1i}가 서로 독립이면 $E(u_{2i}|u_{1i} > -x_{1i}\beta) = E(u_{2i}) = 0$이 성립한다. 하지만 독립이 아니면 $E(u_{2i}|u_{1i} > -x_{1i}\beta) \neq 0$이 되고, 따라서 식 (2.2)의 OLS 추정량은 $E(u_{2i}|u_{1i} > -x_{1i}\beta)$ 부분을 무시하게 되어 그만큼 선택편의가 발생한다.

Heckman (1979)은 오차항 u_{1i}와 u_{2i}가 이변량 정규분포라고 가정하고, 식 (2.2)의 모수 γ에 대한 최우추정량을 제시하였다. 표본선택 모형의 로그우도함수를 일반적 형태로 표현하면 다음과 같다.

$$\begin{aligned} \log L &= \sum_{y_{1i}=0} \log\left[\Pr(y_{1i}=0)\right] + \sum_{y_{1i}=1} \log\left[f(y_{2i}|y_{1i}=1)\Pr(y_{1i}=1)\right] \\ &= \sum_{y_{1i}=0} \log\left[\Pr(y_{1i}=0)\right] + \sum_{y_{1i}=1} \log\left[f(y_{2i})\Pr(y_{1i}=1|y_{2i})\right] \\ &= \sum_{y_{1i}=0} \log\left[\Pr(y_{1i}^* \leq 0)\right] + \sum_{y_{1i}=1} \log\left[f(y_{2i})\Pr\left(y_{1i}^* > 0|y_{2i}\right)\right]. \end{aligned} \tag{2.4}$$

이변량 정규분포 가정 하에서 표본선택 모형의 우도함수는 다음과 같다 (Wooldridge, 2002).

$$\log L = \sum_{y_{1i}=0} \log\left(\Phi(a_{1i}^*)\right) + \sum_{y_{1i}=1} \log\left(\phi\left(\frac{z_{2i}^*}{\sigma_2}\right)\right) + \sum_{y_{1i}=1} \log\left(\Phi\left(\frac{\mu_{12i}^*}{\sigma_{12}^*}\right)\right). \tag{2.5}$$

2.2. 이변량 S_U-정규분포 모형

S_U-정규분포는 정규분포를 변형시킨 것이다 (Johnson 1949a, 1949b). 따라서 이미 잘 정립된 정규분포의 성질을 이용함으로써 쉽게 이변량 분포는 물론, 조건부 확률이나 그 기댓값을 도출할 수 있다. 또한 선택 방정식과 본 방정식의 오차항 분포를 S_U-정규분포로 가정한다면 다양한 비대칭성과 초과 첨도의 비정규성을 포착할 수 있는 장점이 있다.

오차항 u_{1i}와 u_{2i}에 대해 다음과 같이 S_U-정규분포를 가정하자.

$$\begin{aligned} u_{1i} &\sim S_U N(0, 1; \lambda_1, \theta_1), \\ u_{2i} &\sim S_U N\left(0, \sigma_2^2; \lambda_2, \theta_2\right), \end{aligned}$$

여기에서 λ와 θ는 각각 왜도(skewness)와 첨도(kurtosis)를 결정짓는 모수로서 $-\infty < \lambda < \infty, \theta > 0$의 범위를 가진다. λ가 양의 값이면 오른쪽으로 치우친(skewed) 분포이고, 음의 값이면 왼쪽으로 치우친 분포가 된다. 또한 θ는 값이 커질수록 첨도가 커진다. $\theta \to 0$이면 S_U-정규분포는 정규분포에 수렴한다. 따라서 정규분포는 S_U-정규분포의 극한분포이다.

표준정규 확률변수와 (z_{1i}와 z_{2i})가 있을 때, 표본선택 모형의 S_U-정규 확률변수 (u_{1i}와 u_{2i})는 다음과 같이 정의된다.

$$\sinh^{-1}\left(s_1 u_{1i} + m_1\right) = \lambda_1 + \theta_1 z_{1i}, \tag{2.6}$$

$$\sinh^{-1}\left(\frac{s_2 u_{2i}}{\sigma_2} + m_2\right) = \lambda_2 + \theta_2 z_{2i}, \tag{2.7}$$

여기서 m과 s는 S_U-정규 확률변수의 1차 및 2차 중심적률(central moment) 다음과 같다.

$$m_1 = (\omega_1)^{\frac{1}{2}} \sinh(\lambda_1), \qquad m_2 = (\omega_2)^{\frac{1}{2}} \sinh(\lambda_2),$$
$$s_1^2 = \frac{1}{2}(\omega_1 - 1)(\omega_1 \cosh(2\lambda_1) + 1), \qquad s_2^2 = \frac{1}{2}(\omega_2 - 1)(\omega_2 \cosh(2\lambda_2) + 1).$$

위 식에서 $\omega_1 = \exp(\theta_1^2)$, $\omega_2 = \exp(\theta_2^2)$이다. S_U-정규 확률변수 중심적률의 자세한 도출과정은 Choi와 Min (2009)에 나와 있다.

로그우도함수를 위해서는 식 (2.4)에서 이변량 S_U-정규분포를 가정하고 $\Pr[y_{1i}^* \le 0]$, $f(y_{2i})$, $\Pr[y_{1i}^* > 0|y_{2i}]$를 도출해야 한다. 우선 다음이 성립한다.

$$\Pr\left[y_{1i}^* \le 0\right] = \Phi\left(b_{1i}^*\right).$$

위 식에서 $b_{1i}^* = [\sinh^{-1}(m_1 - s_1 x_{1i}\beta) - \lambda_1]/\theta_1$이다. 다음으로 식 (2.7)을 다시 쓰면 $z_{2i} = [\sinh^{-1}(s_2(y_{2i} - x_{2i}\gamma)/\sigma_2 + m_2) - \lambda_2]/\theta_2$이기 때문에 $f(y_{2i})$는 다음과 같다.

$$f(y_{2i}) = J_{2i}\phi\left(v_{2i}^*\right),$$

여기서 $J_{2i} = \partial v_{2i}^*/\partial y_{2i}$이고, $v_{2i}^* = [\sinh^{-1}(s_2(y_{2i} - x_{2i}\gamma)/\sigma_2 + m_2) - \lambda_2]/\theta_2$이다. 로그우도함수의 마지막 요소인 조건부 확률 $\Pr[y_{1i}^* > 0|y_{2i}]$을 구하기 위해서는 y_{2i}가 주어졌을 때 y_{1i}^*의 조건부 기댓값과 분산을 도출해야 한다.

$$\sinh^{-1}\left(s_1\left(y_{1i}^* - x_{1i}\beta\right) + m_1\right)\Big|y_{2i} \sim N\left\{\lambda_1 + \rho_{12}\theta_1 v_{2i}^*, \theta_1^2\left(1 - \rho_{12}^2\right)\right\}. \tag{2.8}$$

위 식에서 ρ_{12}는 표준정규 확률변수들 사이의 상관계수로서 $\mathrm{corr}(z_{1i}, z_{2i}) = \rho_{12}$이다. 식 (2.8)을 이용하면 다음을 도출할 수 있다.

$$\Pr\left[y_{1i}^* > 0|y_{2i}\right] = \Phi\left(\frac{\lambda_1 + \rho_{12}\theta_1 v_{2i}^* - \sinh^{-1}(m_1 - s_1 x_{1i}\beta)}{\theta_1\left(1 - \rho_{12}^2\right)^{\frac{1}{2}}}\right).$$

최종적인 로그우도함수는 다음과 같이 정리할 수 있다.

$$\log L = \sum_{y_{1i}=0} \log\left(\Phi\left(b_{1i}^*\right)\right) + \sum_{y_{1i}=1} \log\left(J_{2i}\phi\left(v_{2i}^*\right)\right) + \sum_{y_{1i}=1} \log\left[\Phi\left(\frac{\lambda_1 + \rho_{12}\theta_1 v_{2i}^* - \sinh^{-1}(m_1 - s_1 x_{1i}\beta)}{\theta_1\left(1 - \rho_{12}^2\right)^{\frac{1}{2}}}\right)\right]. \tag{2.9}$$

2.3. 조건부 기댓값과 선택편의

2.1절과 2.2절에서 제시된 로그우도함수를 이용하면 각 분포함수 모형에서 β와 γ를 추정할 수 있다. 또한 선택편의가 존재하는지에 대한 통계적 판단을 위해 상관계수 $\rho_{12} = 0$의 귀무가설에 대한 가설검정을 할 수 있다. 본 절에서는 분포함수에 따라 추정된 모수, y_{2i}의 조건부 기댓값, 그리고 선택편의의 크기가 어떻게 달라지는지 살펴본다.

표본선택 모형에서 종속변수의 조건부 기댓값이 다음과 같이 표현할 수 있다.

$$E\left(y_{2i}|x_{2i}, y_{1i}^* > 0\right) = x_{2i}\gamma + E\left(u_{2i}|u_{1i} > -x_{1i}\beta\right). \tag{2.10}$$

각 분포별로 식 (2.10)을 도출해보자. 우선 정규분포 모형에서는 오차항 u_{2i}와 u_{1i}가 다음과 같은 관계임을 알 수 있다.

$$u_{2i} = \sigma_{12} u_{1i} + \xi_i. \tag{2.11}$$

위 식에서 ξ_i와 u_{1i}가 서로 독립이면 $\xi_i \sim N(0, (1-\rho_{12})\sigma_2^2)$의 분포를 가지는 정규 확률변수임을 쉽게 증명할 수 있다. 식 (2.11)을 식 (2.10)에 대입하면, 정규분포 모형에서 조건부 기댓값은 다음과 같다 (Heckman, 1979).

$$x_{2i}\gamma + E\left(u_{2i} | u_{1i} > -x_{1i}\beta\right) = x_{2i}\gamma + \sigma_{12}\frac{\phi(x_{1i}\beta)}{\Phi(x_{1i}\beta)}. \tag{2.12}$$

다음으로 이변량 S_U-정규분포 모형에서 조건부 기댓값을 구하기 위해서는 먼저 식 (2.10)의 $E(u_{2i}|u_{1i} > -x_{1i}\beta)$을 구해야 한다. 우선 u_{1i}가 주어졌을 때 u_{2i}의 조건부 기댓값 $E(u_{2i}|u_{1i})$은 다음과 같다.

$$E(u_{2i}|u_{1i}) = \sigma_2 \left(\exp\left(\frac{1}{2}\theta_{21}^{*2}\right) \sinh\left(\lambda_{21}^*\right) - m_2 \right) / s_2, \tag{2.13}$$

여기에서 $\theta_{21}^* = \theta_2(1-\rho_{12}^2)^{1/2}$, $\lambda_{21}^* = \lambda_2 + \rho\theta_2(\sinh^{-1}(s_1 u_{1i} + m_1) - \lambda_1)/\theta_1$이다. 식 (2.13)을 이용하면 $u_{1i} > -x_{1i}\beta$의 조건이 주어졌을 때의 기댓값을 쉽게 도출할 수 있다. 이는 다음과 같다.

$$E\left(u_{2i} | u_{1i} > -x_{1i}\beta\right) = \sigma_2 \left(\exp\left(\frac{1}{2}\theta_{21}^*\right) M_{21} - m_2 \right) / s_2, \tag{2.14}$$

여기에서 $M_{21} = E[\sinh(\lambda_{21}^*)|u_{1i} > -x_{1i}\beta]$으로서 다음과 같다.

$$\begin{aligned} M_{21} &= E\left[\sinh\left(\lambda_2 + \rho\theta_2 z_{1i}\right) | z_{1i} > c_1^*\right] \\ &= \frac{\exp\left(\frac{1}{2}\rho^2\theta_2^2 + \lambda_2\right)\Phi\left(\rho_{12}\theta_2 - c_{1i}^*\right) - \exp\left(\frac{1}{2}\rho_{12}^2\theta_2^2 - \lambda_2\right)\Phi\left(-\rho_{12}\theta_2 - c_{1i}^*\right)}{2\left(1 - \Phi\left(c_{1i}^*\right)\right)}, \end{aligned} \tag{2.15}$$

여기에서 $c_{1i}^* = (\sinh^{-1}(-x_{1i}\beta s_1 + m_1) - \lambda_1)/\theta_1$이다.

이상 각 분포함수에서 조건부 기댓값을 구했기 때문에 선택편의의 크기 역시 쉽게 구할 수 있다. 먼저 정규분포 모형에서 선택편의는 다음과 같다.

$$E(y_{2i}|y_{1i} > 0) - E(y_{2i}) = x_{2i}\gamma + \sigma_{12}\frac{\phi(x_{1i}\beta)}{\Phi(x_{1i}\beta)} - x_{2i}\gamma = \sigma_{12}\frac{\phi(x_{1i}\beta)}{\Phi(x_{1i}\beta)}. \tag{2.16}$$

다음으로 S_U-정규분포 모형에서 선택편의는 다음과 같다.

$$\begin{aligned} E\left(u_{2i}|y_{1i} > 0\right) - E(y_{2i}) &= x_{2i}\gamma + \sigma_2\left(\exp\left(\frac{1}{2}\theta_{21}^*\right)M_{21} - m_2\right)/s_2 - x_{2i}\gamma \\ &= \sigma_2\left(\exp\left(\frac{1}{2}\theta_{21}^*\right)M_{21} - m_2\right)/s_2. \end{aligned} \tag{2.17}$$

표 1: DGP I(이변량 정규분포) 하에서의 모수 추정결과

모수	실제값	정규분포 모형		S_U-정규분포 모형	
		n = 1000	n = 2000	n = 1000	n = 2000
γ_0	−5.000	−5.001 (0.621)	−4.995 (0.444)	−4.675 (0.858)	−4.696 (0.743)
γ_1	1.000	1.001 (0.078)	1.000 (0.055)	0.961 (0.106)	0.963 (0.092)
σ_2	2.000	2.000 (0.066)	2.000 (0.048)	1.988 (0.070)	1.990 (0.052)
ρ_{12}	0.500	0.497 (0.140)	0.498 (0.101)	0.429 (0.198)	0.436 (0.167)
로그우도 함수		−1220.907	−2444.961	−1224.037	−2451.624

주: 괄호 안의 값은 RMSE임

표 2: DGP II(이변량 S_U-정규분포) 하에서의 모수 추정결과

모수	실제값	정규분포 모형		S_U-정규분포 모형	
		n = 1000	n = 2000	n = 1000	n = 2000
γ_0	−5.000	−5.482 (1.623)	−5.603 (1.334)	−4.993 (0.260)	−5.001 (0.184)
γ_1	1.000	1.054 (0.202)	1.069 (0.164)	0.999 (0.031)	1.000 (0.022)
σ_2	2.000	2.206 (0.422)	2.230 (0.360)	2.009 (0.220)	2.002 (0.151)
ρ_{12}	0.500	0.618 (0.257)	0.642 (0.228)	0.497 (0.098)	0.499 (0.070)
로그우도 함수		−1244.753	−2509.365	−974.476	−1953.708

주: 괄호 안의 값은 RMSE임

3. 시뮬레이션

본 장에서는 시뮬레이션을 통해 정규분포 모형과 S_U-정규분포 모형에서 구한 추정량의 불편성(unbiasedness)와 일치성(consistency)을 비교분석 한다. 시뮬레이션에서는 선택 방정식과 본 방정식 모두 상수항 외에 1개의 설명변수만 있는 가장 간단한 모형을 설정하였다.

$$y_{1i}^* = \beta_0 + \beta_1 x_{1i} + u_{1i} : \text{ 선택 방정식}$$
$$y_{2i} = \gamma_0 + \gamma_1 x_{2i} + u_{2i} : \text{ 본 방정식}$$

여기에서 $\beta_0 = \gamma_0 = -5$, 그리고 $\beta_1 = \gamma_1 = 1$로 정하였다. 설명변수 x_{1i}와 x_{2i}는 균등분포 $U(0, 10)$에서 무작위로 추출하였고 $x_{1i} = x_{2i}$로 설정하였다. 선택 방정식에서 표본크기(n)는 1,000과 2,000의 두 가지를 시도하였다. 앞에서 가정한 모수값 하에서 $y_{1i}^* > 0$인 표본 수는 전체 표본의 약 50%가 된다.

오차항 u_{1i}과 u_{2i}에 대해서는 다음의 두 가지 데이터 발생과정(DGP; data generating process)을 가정하였다.

$$\text{DGP I: 이변량 정규 확률분포 } u_{1i} \sim N(0,1),\ u_{2i} \sim N\left(0, 2^2\right)$$
$$\text{DGP II: 이변량 } S_U\text{-정규 확률분포 } u_{1i} \sim SuN(0,1;1,1),\ u_{2i} \sim SuN\left(0, 2^2; 1, 1\right)$$

여기에서 S_U-정규분포는 $\lambda = 1$로서 양의 왜도를 가지고 있고, θ 역시 1로서 정규분포보다 꼬리가 훨씬 두꺼운 분포를 가정하였다. 두 DGP 모두 $\text{corr}(u_{1i}, u_{2i}) = 0.5$로 설정하였다.

표 1과 표 2는 각각 DGP I과 DGP II에 의한 시뮬레이션 데이터에 대해 정규분포 모형과 S_U-정규분포 모형으로 추정한 결과를 비교해서 보여주고 있다. 각 시뮬레이션 시행에서 추정치를 얻고, 동일한 방식의 시뮬레이션을 1,000번 반복 시행하여 그 추정치의 평균값과 RMSE(root mean squared error)를 계산한 것이 표에 나와 있다.

우선 DGP가 이변량 정규분포인 경우가 표 1에 나와 있다. DGP가 정규분포이기 때문에 정규분포 모형의 추정성과가 더 우수할 것으로 예상할 수 있다. 문제는 S_U-정규분포 모형의 추정성과가 정규분

포 모형에 비해 얼마만큼 뒤지지 않느냐이다. 표를 보면 예상대로 정규분포 모형의 모수 추정치는 모두 불편추정치임을 인정할 수 있다. 본 방정식의 모수(γ_1, γ_2)뿐만 아니라 오차항의 표준편차(σ_2) 및 상관계수(ρ_{12}) 추정치 모두 시뮬레이션 평균값이 실제값과 거의 차이가 없는 것을 알 수 있다. 추정량의 표준오차 개념인 RMSE는 예상대로 표본크기가 1,000에서 2,000으로 커짐에 따라 값이 작아지는 것을 확인할 수 있다. 이러한 정규분포 모형의 추정성과에 비교해서 S_U-정규분포 모형의 추정성과를 살펴보면, 전체적으로 정규분포 모형에 비해 불편성(unbiasedness) 및 RMSE 모두에서 뒤지는 것을 알 수 있다. 모수 추정의 불편성에 있어서는 σ_2를 제외하고는 정규분포 모형과 같은 확실한 불편성을 보이지는 않는 것으로 나타났다. 다만 표본크기가 커짐에 따라 추정치 평균이 모든 모수에서 실제치에 더 근접해가는 것을 알 수 있다. 또한 RMSE를 비교해보면 모든 경우에 S_U-정규분포 모형이 정규분포 모형에 비해 값이 더 크기는 하지만, 그 차이가 나중에 살펴볼 DGP II 하에서 두 모형의 성과 차이에 비해서는 훨씬 작은 편이다. 즉 S_U-정규분포 모형의 RMSE는 정규분포 모형에 비해 최저 1.1배에서 최대 1.7배 수준이나, 나중에 표 2에서 살펴볼 DGP II 하에서는 정규분포 모형의 RMSE가 S_U-정규분포 모형에 비해 최대 7.5배나 더 크다는 점에서 S_U-정규분포 모형의 추정성과가 상대적으로 양호한 것을 알 수 있다. 표의 마지막 행에 나와 있는 로그우도 값에서도 두 모형 간에 차이가 크지 않음을 확인할 수 있다.

이번에는 DGP가 이변량 S_U-정규분포 경우인 표 2의 결과를 살펴보자. 앞에서와 마찬가지 논리로 DGP가 S_U-정규분포이기 때문에 S_U-정규분포 모형의 추정성과가 더 우수할 것으로 짐작할 수 있으며, 문제는 정규분포 모형이 S_U-정규분포 모형에 비해 어느 정도 성과를 보이느냐이다. 표를 보면 S_U-정규분포 모형의 모수 추정치는 모두 불편추정치임을 인정할 수 있다. 모든 모수($\gamma_1, \gamma_2, \sigma_2, \rho_{12}$)에 대해 시뮬레이션 평균값이 실제값과 거의 차이가 없다. 이에 반해 정규분포 모형의 추정성과를 보면, 전체적으로 불편추정량으로 보기 어려운 것을 알 수 있다. 또한 앞의 표 1의 결과와 달리 표 2에서는 표본크기가 1,000에서 2,000으로 커짐에 따라 정규분포 모형의 추정치가 4개 모수 모두에 있어서 실제값에 가까워지는 것이 아니라 오히려 멀어지는 것으로 나타났다. 뿐만 아니라 RMSE를 비교해보면 표 1 설명 과정에서도 언급했듯이 DGP I 하에서는 S_U-정규분포 모형이 정규분포 모형에 비해 값이 더 크기는 하지만 그 차이가 그다지 크지 않은 것에 반해, 표 2 DGP II 하에서 정규분포 모형의 RMSE는 S_U-정규분포 모형에 비해 최대 7.5배나 더 큰 것으로 나타났다. 마지막으로 표의 마지막 행에 나와 있는 로그우도 값에서도 DGP I에서와 달리 두 모형 간에 차이가 상당히 큰 것을 확인할 수 있다.

이상 표본선택 모형의 본 방정식 모수 추정치과 그 RMSE를 비교했다. 그런데 표본선택 모형에서는 모수 자체의 불편성 및 일치성도 중요하지만, 조건부 기댓값이나 한계효과(marginal effect)의 불편성 및 일치성이 더 중요한 의미를 지닌다. 이하에서는 식 (2.12) 및 (2.14)에 제시된 정규분포 모형 및 S_U-정규분포 모형의 조건부 기댓값 추정결과에 대해 살펴보자. 아래의 표 3과 표 4는 앞의 표 1과 표 2에서와 동일한 DGP I(이변량 정규분포) 및 DGP II(이변량 S_U-정규분포) 데이터에 대해 조건부 기댓값을 추정한 결과를 모형별 및 표본크기별로 대비해서 보여주고 있다. 그런데 표본선택 모형에서는 일반적인 선형회귀 모형과 달리 설명변수의 값에 따라 조건부 기댓값이 달라진다. 본 시뮬레이션에서 설명변수의 범위를 $U(0, 10)$으로 했기 때문에 여기에서는 $x_{2i} = 3, 5, 7$의 세 가지 값에서 측정한 조건부 기댓값을 기준으로 추정성과를 비교했다.

표 3에서는 DGP가 정규분포인 상황 하에서, 정규분포 모형과 S_U-정규분포 모형으로 추정한 조건부 기댓값을 비교하고 있다. DGP가 정규분포이기 때문에 정규분포 모형의 추정성과가 더 우수할 것으로 예상할 수 있으며, 문제는 S_U-정규분포 모형의 추정성과가 정규분포 모형에 비해 얼마만큼 뒤지지 않느냐이다. 표를 보면 예상대로 정규분포 모형의 추정치는 모든 x값에서 불편추정치임을 인정할 수 있다. $x = 3$에서의 추정치가 실제값보다 다소 크지만, 과대평가 정도는 2.7%(표본크기 1,000인 경우) 수준에 불과하다. 이러한 정규분포 모형의 추정성과에 비교해서 S_U-정규분포 모형의 추정성과를 살

표 3: DGP I(이변량 정규분포) 하에서의 조건부 기댓값 추정결과

x	실제값	정규분포 모형		S_U-정규분포 모형	
		n = 1000	n = 2000	n = 1000	n = 2000
x = 3	0.373	0.383 (0.476)	0.380 (0.331)	0.374 (0.680)	0.362 (0.505)
x = 5	0.798	0.798 (0.154)	0.801 (0.108)	0.831 (0.175)	0.832 (0.130)
x = 7	2.055	2.059 (0.111)	2.059 (0.080)	2.092 (0.137)	2.089 (0.109)

주: 괄호 안의 값은 RMSE임

표 4: DGP II(이변량 S_U-정규분포) 하에서의 조건부 기댓값 추정결과

x	실제값	정규분포 모형		S_U-정규분포 모형	
		n = 1000	n = 2000	n = 1000	n = 2000
x = 3	1.317	0.794 (0.890)	0.839 (0.699)	1.431 (0.813)	1.361 (0.533)
x = 5	1.104	0.897 (0.332)	0.892 (0.278)	1.107 (0.212)	1.103 (0.149)
x = 7	2.002	2.021 (0.137)	1.998 (0.094)	2.006 (0.088)	2.003 (0.062)

주: 괄호 안의 값은 RMSE임

펴보면, 전체적으로 정규분포 모형에 비해서는 뒤지지만, 차이가 크지는 않다고 말할 수 있다. S_U-정규분포 모형의 조건부 기댓값 추정치를 보면, $x = 3, 7$의 경우에는 정규분포 모형과 마찬가지로 불편성을 보이는 것으로 간주할 수 있다. 다만, $x = 5$에서는 추정치 평균이 실제값보다 9% 가량 과대평가된 것으로 나타났다. 한편 RMSE를 비교해보면 모든 경우에 S_U-정규분포 모형이 정규분포 모형에 비해 값이 더 크기는 하지만, 그 차이가 그다지 큰 편은 아니다. 즉 S_U-정규분포 모형의 RMSE는 정규분포 모형에 비해 최대 1.5배 수준이다.

마지막으로 DGP가 이변량 S_U-정규분포 경우일 때 조건부 기댓값 추정결과를 정리해 놓은 표 4를 살펴보자. DGP가 S_U-정규분포이기 때문에 S_U-정규분포 모형의 추정성과가 더 우수할 것으로 짐작할 수 있다. 표를 보면 S_U-정규분포 모형의 모수 추정치는 대체로 불편추정치임을 인정할 수 있다. $x = 3$일 때 추정치가 실제보다 약간 과대평가되었지만 그 차이는 크지 않고, 이를 제외하면 시뮬레이션 평균값이 실제값과 거의 차이가 없다. 이에 반해 정규분포 모형의 추정성과를 보면, 전체적으로 불편추정량으로 보기 어려운 것을 알 수 있다. 특히 $x = 3$일 때 추정치가 실제값을 대폭 과소평가하는 것을 알 수 있다. RMSE를 비교해보면, 주어진 표본크기와 x값에서 정규분포 모형이 S_U-정규분포 모형에 비해 항상 RMSE 값이 더 큰 것을 확인할 수 있으며, 그 차이는 최대 1.9배 수준이다.

4. 실증분석

4.1. 실증분석 모형 및 데이터 설명

본 장에서는 실제 데이터를 이용한 추정을 시도하여 정규분포 모형과 S_U-정규분포 모형의 추정결과를 비교한다. 실증분석 예로는 자동차 보유 여부(ownership)와 자동차 보유 가구의 차량 유지비(expenditure)를 각각 선택 방정식과 본 방정식의 종속변수로 두었다. 표본선택 모형은 다음과 같다.

$$\begin{aligned} \text{선택 방정식:} \quad & \text{own}_i^* = x_{1i}\beta + u_{1i} \\ \text{본 방정식:} \quad & \text{exp}_i = x_{2i}\gamma + u_{2i} \end{aligned} \tag{4.1}$$

선택 방정식에서 own_i^*는 관찰되지 않는 종속변수이다. 만약 $\text{own}_i^* > 0$이면 관찰된 종속변수는 $\text{own}_i = 1$이고, $\text{own}_i^* \le 0$이면 $\text{own}_i = 0$이 된다. $\text{own}_i = 1$이 자동차 보유를 의미하고, 자동차가 없으면 $\text{own}_i = 0$이다. exp_i는 자동차 보유 가구의 한 달 자동차 유지비이다. 일반적으로 자동차가 필요한 가구일수록

표 5: 자동차 보유 비율과 평균 유지비용

전체 가구	자동차 보유 가구 (비율)	월 평균 유지비용
5,116 가구	3,056 가구 (59.7%)	28.3만원

자동차를 구매할 가능성이 높을 것이다. 또한 자동차가 필요한 가구일수록 자동차를 더 많이 운행하고 이에 따라 유지비가 더 높아질 가능성이 있다. 즉 자동차 보유 의사결정이 자동차 유지비와 연결되어 있는 것으로 볼 수 있다. 자동차 보유 여부와 자동차 사용 정도에 대한 의사결정이 서로 동시적으로 결정된다면, 선택 방정식의 오차항과 본 방정식의 오차항은 상관성을 지니게 된다. 즉 $\rho_{12} \neq 0$이다. 이런 상황에서 (자동차 보유 가구의) 자동차 유지비를 분석하고자 할 때, 식 (4.1)로 주어진 표본선택 모형이 정확한 분석을 가능하게 한다.

실증분석 데이터는 2008년 한국노동패널(KLIPS; Korea Labor and Income Panel Survey) 가구용 데이터를 사용하였다. 가구용 데이터에는 각 가구의 자동차 보유 여부와 자동차 유지비가 변수로 들어있다. 표 5에 자동차 보유가구 비율과 자동차 평균 유지비에 대한 요약통계량이 나와 있다. 전체 가구 중 약 60%인 3,056 가구가 자동차를 보유하고 있고, 자동차 보유가구의 월평균 유지비용은 28.3만원인 것으로 나타났다.

표본선택 모형에 들어가는 설명변수는 Klaauw와 Koning (2003)와 Vythoulkas (2007)의 연구를 참고하여 자동차 보유와 유지비에 영향을 미칠 것으로 예상되는 특성변수들을 선택하였다. 사용된 설명변수는 가구원 수(fsize), 가구주 연령(hage) 및 연령 제곱값(sqage), 성별(gender), 가구주 교육수준(중졸이하, 고졸, 대졸이상), 현재 고용상태(취업과 미취업), 소득(lincome), 거주지역(서울·광역시와 기타지역) 등이다. 선택 방정식과 본 방정식에서 동일한 설명변수를 사용하였으며, 자동차 유지비와 소득 변수에는 로그를 취하였다.

4.2. 추정결과

식 (4.1)의 표본선택 모형을 정규분포 모형과 S_U-정규분포 모형으로 추정한 결과가 표 6에 나와 있다. 선택편의를 고려하지 않은 결과와 비교하기 위해 본 방정식에 대한 OLS 추정 결과도 함께 살펴보았다. OLS 추정량은 선택 방정식과 본 방정식의 오차항의 상관관계가 0이고, 본 방정식의 오차항을 정규분포로 가정하였을 때 얻은 결과와 동일하다. 하지만 선택편의가 있는 경우 OLS 추정량은 일치추정량이 되지 못한다.

표의 추정결과를 변수별로 보면, 우선 가구원 수(fsize) 변수는 정규분포 모형의 추정 계수(0.037)가 S_U-정규분포 모형(0.022)보다 70% 가량 커 두 모형 간에 차이가 큰 것으로 나타났다. 특히 주목할 만한 변수는 성별(gender) 변수로서 S_U-정규분포 모형에서는 계수 부호가 양이고 10% 수준에서 유의한 반면, 정규분포 모형에서는 음의 값이고 10% 수준에서 유의하지 않는 등 모형 간에 큰 차이를 보이고 있다. 직관적으로 생각해보면, 남자 가구주가 여자 가구주보다 자동차 유지비용 지출이 평균적으로 더 많은 것으로 나타난 S_U-정규분포 모형의 추정 결과가 직관과 더 일치한다고 말할 수 있다. 소득(lincome) 변수의 추정치에서도 두 모형 간 차이가 있다. S_U-정규분포 모형의 추정 계수(0.272)가 정규분포 모형(0.163) 보다 훨씬 크다. 선택편의가 없다고 가정한 OLS 모형에서 소득 변수 추정치는 0.139로 선택편의를 가정한 두 모형의 추정치보다 작다는 것을 알 수 있다. 한편, S_U-정규분포 모형에서 제시된 λ_1(0.349)와 λ_2(−0.412) 추정치로 판단하면, 오차항의 분포는 비대칭(skewed) 분포이다. 또한 θ_1(0.939)과 θ_2(0.690) 추정치가 0보다 훨씬 큰 값이기 때문에 정규분포 가정이 적절치 않는 것으로 보인다.

표 6: 실증분석 모형 추정결과

	변수	OLS 추정	정규분포 모형	S_U-정규분포 모형
본 방정식	fsize	0.024***	0.037***	0.022***
		(0.009)	(0.009)	(0.008)
	hage	0.017***	0.020***	0.017***
		(0.006)	(0.006)	(0.005)
	sqage	−0.016***	−0.021***	−0.017***
		(0.006)	(0.006)	(0.005)
	gender	−0.024	−0.010	0.050*
		(0.032)	(0.033)	(0.030)
	dum_edu2	0.034	0.083***	0.065***
		(0.031)	(0.033)	(0.029)
	dum_edu3	0.151***	0.227***	0.200***
		(0.033)	(0.038)	(0.032)
	employed	−0.098	−0.068	−0.008
		(0.060)	(0.060)	(0.060)
	lincome	0.139***	0.163***	0.272***
		(0.012)	(0.013)	(0.022)
	region	−0.071***	−0.088***	−0.095***
		(0.019)	(0.020)	(0.018)
	상수	1.627***	1.122***	0.266
		(0.161)	(0.206)	(0.233)
선택 방정식	fsize		0.202***	0.092***
			(0.018)	(0.019)
	hage		0.054***	0.032***
			(0.010)	(0.009)
	sqage		−0.060***	−0.035***
			(0.010)	(0.009)
	gender		0.388***	0.257***
			(0.055)	(0.052)
	dum_edu2		0.495***	0.288***
			(0.054)	(0.054)
	dum_edu3		0.987***	0.550***
			(0.063)	(0.088)
	employed		0.158*	0.134**
			(0.084)	(0.068)
	lincome		0.289***	0.366***
			(0.024)	(0.044)
	region		−0.288***	−0.204***
			(0.041)	(0.037)
	상수		−4.494***	−4.051***
			(0.307)	(0.528)
ρ			0.286	0.464
σ_2			0.547	0.626
λ_1				0.349
θ_1				0.939
λ_2				−0.412
θ_2				0.690
$\log L$			−4,843.37	−4,696.86
관측치		3,020	선택 방정식: 5,079 본 방정식: 3,020	

주: 1. ***, **, *은 각각 1%, 5%, 10% 수준에서 통계적으로 유의함을 의미함
2. 괄호 안의 값은 추정치의 표준오차임.

표 7: LR 검정: 선택편의 검정

	정규분포 모형	S_U-정규분포 모형
$\rho_{12} = 0$	−4848.35	−4706.52
$\rho_{12} \neq 0$	−4843.37	−4696.86
검정통계량	9.96	19.32
검정결과	H_0 기각	H_0 기각

주: $\chi^2(1)$의 1% 임계치는 6.634임

표 8: LR 검정: 정규성 검정

정규분포 모형	S_U-정규분포 모형	검정통계량	검정결과
−4843.37	−4696.86	293.02	H_0 기각

주: $\chi^2(3)$의 1% 임계치는 11.34임

이번에는 로그우도함수 값을 이용하여 선택편의가 있는지를 검정하였다. LR(likelihood ratio) 검정통계량은 다음과 같다.

$$\text{LR} = -2\left(\ln L_{\rho_{12}=0} - \ln L_{\rho_{12}\neq 0}\right) \sim \chi^2(1).$$

표 7의 선택편의 검정 결과를 보면, 정규분포 모형과 S_U-정규분포 모형 어느 것을 가정하든 모두 선택편의가 없다는 귀무가설이 기각되는 것으로 나타났다. 즉 자동차 보유 방정식과 자동차 유지비 방정식이 서로 상관관계를 가지고 있다고 가정할 필요가 있다는 것을 의미한다.

다음으로 정규분포 가정에 대한 검정을 시도하였다. 검정의 귀무가설과 대립가설은 다음과 같다.

$$H_0: \text{ 오차항 } u_{1i}\text{와 } u_{2i}\text{는 이변량 정규분포를 따른다.}$$
$$H_1: \text{ 오차항 } u_{1i}\text{와 } u_{2i}\text{는 } S_U\text{-정규분포를 따른다.}$$

S_U-정규분포에서 $\theta \to 0$이면 정규분포에 수렴한다. 따라서 $\theta_1 \to 0$이고, $\theta_2 \to 0$이면 S_U-정규분포 모형의 추정치는 정규분포 모형의 추정치와 비슷하고 로그우도함수 값도 서로 비슷할 것으로 예상할 수 있다. 따라서 다음과 같이 LR 검정통계량을 사용할 수 있다.

$$\text{LR} = -2\left(\ln L_{Normal} - \ln L_{SuN}\right)$$

정규분포가 S_U-정규분포의 '극한(limiting)' 분포이기 때문에 위 검정통계량이 $\chi^2(2)$를 따른다고 말할 수 없다. Bollerslev (1987)와 최필선과 민인식 (2009)은 부트스트랩(bootstrapping)으로 검정통계량의 임계치를 찾는 방법을 제시하고 있다. 선행연구의 결과를 응용하면 위 검정통계량의 임계치는 $\chi^2(2)$에 의한 임계치보다 큰 값이라고 말할 수 있다. 따라서 본 연구에서는 보수적인 가설검정을 위해 $\chi^2(3)$에서 구한 임계치와 LR 검정통계량을 비교하여 기각여부를 결정한다. 아래 표에서 제시되었듯이 LR 검정통계량은 293.02로 임계치(11.34)를 대폭 상회하는 것으로 나타났다. 따라서 오차항이 이변량 정규분포를 따른다는 귀무가설을 기각할 수 있다. 이를 받아들이면 정규분포 모형으로 구한 최우추정량은 일치추정량이 되지 못한다.

5. 요약 및 결론

표본선택 모형에서 오차항 분포를 일반적으로 이변량 정규분포로 가정하지만, 이 가정이 오차항의 실제 분포를 과도하게 제약할 가능성이 있다. 본 연구는 표본선택 모형의 오차항 분포로 이변량 S_U-정규분포를 도입했다. S_U-정규분포는 분포의 비대칭성과 초과첨도를 허용한다는 측면에서 정규분포보

다 훨씬 유연하면서, 동시에 정규분포를 극한분포의 형태로 포함하고 있다. 또한 정규분포처럼 다변량 분포함수가 존재하기 때문에 표본선택 모형과 같은 다변량 모형에서도 활용할 수 있다.

본 논문은 우선 S_U-정규분포를 이용한 표본선택 모형을 제시하고 로그우도 함수와 조건부 기댓값을 도출했다. 또한 시뮬레이션을 통해 S_U-정규분포 모형과 정규분포 모형의 추정성과를 비교했다. 시뮬레이션은 표본선택 모형의 오차항을 정규분포와 S_U-정규분포 등 두 가지로 DGP를 가정한 다음, 각 DGP 하에서 정규분포 모형과 S_U-정규분포 모형을 사용하여 최우추정법으로 추정했다. 추정결과를 통해 우리는 표본선택 모형에 있어서 S_U-정규분포 모형의 유용성을 확인할 수 있었다. 즉 DGP가 이변량 정규분포인 경우에는 불편성 및 RMSE 등에 있어서 정규분포 모형의 추정성과가 더 우수하기는 하지만, S_U-정규분포 모형도 이에 크게 뒤지지 않는 것으로 나타났다. 이에 반해 DGP가 이변량 S_U-정규분포 경우에는 정규분포 모형의 추정성과가 상대적으로 크게 악화되었는데, 특히 RMSE에 있어서 정규분포 모형이 S_U-정규분포 모형에 비해 최대 7.5배나 더 크게 나왔다. 이러한 결과는 모수뿐만 아니라 조건부 기댓값의 추정에서도 비슷한 것으로 나타났다.

한편 S_U-정규분포 모형을 실제 데이터에 적용하여 자동차 보유 가구들의 자동차 유지비를 분석한 결과, 추정 계수의 크기, 부호, 유의성 등에서 정규분포 모형과 크게 달라지는 경우가 있는 것으로 나타났다. 또한 LR 검정을 통해 정규성을 검정한 결과, 오차항이 이변량 정규분포를 따른다는 귀무가설이 기각되었다. 이를 받아들이면 정규분포 모형으로 구한 최우추정량은 일치추정량이 되지 못하지 때문에 S_U-정규분포 모형을 사용해야 한다는 것을 의미한다.

부록: S_U-정규분포 모형에서 조건부 기댓값 도출

이변량 S_U-정규분포에서 조건부 기댓값을 도출하기 위해서는 먼저 단일변량 S_U-정규 확률변수에서 조건부 기대값 공식을 알아야 한다. S_U-정규 확률변수 y는 다음과 같이 정규 확률변수의 함수로 표현할 수 있다.

$$\sinh^{-1}(y) = \lambda + \theta z.$$

위 식에서 z는 표준정규 확률변수이다. $y > a$의 조건 하에서 기댓값은 다음과 같다.

$$E(y|y > a) = E\left[\sinh(\lambda + \theta z)|z \geq a^*\right]. \tag{A.1}$$

위 식에서 $a^* = (\sinh^{-1}(a) - \lambda)/\theta$이다. 식 (A.1)을 적분을 이용하여 계산하면 다음과 같다.

$$\begin{aligned} E\left[\sinh(\lambda + \theta z)|z \geq a^*\right] &= \frac{1}{\Pr(z \geq a^*)} \int_{a^*}^{\infty} \sinh(\lambda + \theta z) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) dz \\ &= \frac{\exp\left(\frac{1}{2}\theta^2 + \lambda\right)\Phi\left(\theta - a^*\right) - \exp\left(\frac{1}{2}\theta^2 - \lambda\right)\Phi\left(-\theta - a^*\right)}{2\left(1 - \Phi(a^*)\right)}. \end{aligned} \tag{A.2}$$

위 식의 두 번째 등호의 구체적인 도출과정은 Choi와 Min (2011)을 참고하라. 식 (A.2)의 결과를 이용하여 오차항이 이변량 S_U-정규분포일 때 조건부 기댓값을 계산할 수 있다. 조건부 기댓값을 다시 쓰면 다음과 같다.

$$E\left(y_{2i}|x_{2i}, u_{1i} > -x_{1i}\beta\right) = x_{2i}\gamma + E\left(u_{2i}|u_{1i} > -x_{1i}\beta\right).$$

본문의 식 (2.13)을 이용하면

$$E(u_{2i}) = E\left(E(u_{2i}|u_{1i})\right) = \sigma_2\left(\exp\left(\frac{1}{2}\theta_{21}^{*2}\right)E\left[\sinh(\lambda_{21}^*)\right] - m_2\right)/s_2. \tag{A.3}$$

이제 $M_{21} = E[\sinh(\lambda_{21}^*)|u_{1i} > -x_{1i}\beta]$로 정의하고 식 (A.3)을 이용하면 본문 식 (2.14)의 조건부 기댓값이 도출된다. 마지막으로 M_{21}은 다음과 같이 풀어서 쓸 수 있다.

$$\begin{aligned} M_{21} &= E\left[\sinh(\lambda_{21}^*)|u_{1i} > -x_{1i}\beta\right] \\ &= E\left[\sinh(\lambda_2 + \rho_{12}\theta_2 z_{1i})|z_{1i} > c_1^*\right] \\ &= \frac{\exp\left(\frac{1}{2}\rho_{12}^2\theta_2^2 + \lambda_2\right)\Phi\left(\rho_{12}\theta_2 - c_{1i}^*\right) - \exp\left(\frac{1}{2}\rho_{12}^2\theta_2^2 - \lambda_2\right)\Phi(-\rho_{12}\theta_2 - c_{1i}^*)}{2(1 - \Phi(c_{1i}^*))}. \end{aligned}$$

위 식에서 $c_{1i}^* = (\sinh^{-1}(-x_{1i}\beta s_1 + m_1) - \lambda_1)/\theta_1$이다. 식 (A.4)의 마지막 결과는 식 (A.2)에서 주어진 단일변량일 때 조건부 기댓값 공식을 그대로 이용한 것이다.

참고 문헌

최필선, 민인식 (2009). Further applications of Johnson's S_U-normal distribution to various regression models, *Communications of the Korean Statistical Society*, **15**, 1–11.

Bollerslev, T. (1987). A conditionally heteroskedastic time series model for speculative prices and rates of return, *Review of Economics and Statistics*, **69**, 542–547.

Choi, P. and Min, I. (2009). Estimating endogenous switching regression model with a flexible parametric distribution function: Application to Korean housing demand, *Applied Economics*, **41**, 3045–3055.

Choi, P. and Min, I. (2011). A comparison of conditional and unconditional approaches in Value-at-Risk estimation, *Japanese Economic Review*, **62**, 99–115.

Heckman, J. (1979). Sample selection bias as a specification error, *Econometrica*, **47**, 153-161.

Johnson, N. (1949a). Systems of frequency curves generated by method of translation, *Biometrika*, **36**, 149–176.

Johnson, N. (1949b). Bivariate distributions based on simple translation systems, *Biometrika*, **36**, 297–304.

Klaauw, B. and R. Koning (2003). Testing the normality assumption in the sample selection model with an application to travel demand, *Journal of Business and Economic Statistics*, **21**, 31–42.

Manski, C. (1989). Anatomy of the selection problem, *Journal of Human Resources*, **24**, 343–360.

Vythoulkas, P. (2007). Car ownership and household transport expenditure in Greece, Working paper, *National Technical University of Athens*.

Wooldridge, J. (2002). *Econometric Analysis of Cross Section and Panel Data*, MIT Press, London, England.

2011년 10월 6일 접수; 2012년 2월 7일 수정; 2012년 3월 19일 채택

An Alternative Parametric Estimation of Sample Selection Model: An Application to Car Ownership and Car Expense

Phil Sun Choi[a], In Sik Min[1,b]

[a]Department of International Trade, Konkuk University
[b]Department of Economics, Kyung Hee University

Abstract

In a parametric sample selection model, the distribution assumption is critical to obtain consistent estimates. Conventionally, the normality assumption has been adopted for both error terms in selection and main equations of the model. The normality assumption, however, may excessively restrict the true underlying distribution of the model. This study introduces the S_U-normal distribution into the error distribution of a sample selection model. The S_U-normal distribution can accommodate a wide range of skewness and kurtosis compared to the normal distribution. It also includes the normal distribution as a limiting distribution. Moreover, the S_U-normal distribution can be easily extended to multivariate dimensions. We provide the log-likelihood function and expected value formula based on a bivariate S_U-normal distribution in a sample selection model. The results of simulations indicate the S_U-normal model outperforms the normal model for the consistency of estimators. As an empirical application, we provide the sample selection model for car ownership and a car expense relationship.

Keywords: Sample selection model, maximum likelihood estimation, S_U-normal distribution.

This work was supported by the research grant of Kyung Hee University in 2012.

[1] Corresponding author: Associate Professor, Department of Economics, Kyung Hee University, Seoul 130-701, Korea.
E-mail: jjcho@chungbuk.ac.kr