

A Two-Step Multinomial Probit Estimator for Correcting Asymmetric Sample Selection Bias

Insik Min*

Abstract This study proposes a two-step estimator based on a multinomial probit model (MNP-TS) to correct sample selection bias arising in multinomial choice models. The conventional multinomial logit-based correction method relies on a correction term based on a single probability index, which may distort estimation when unobserved factors generate asymmetric bias across alternatives. To address this limitation, this study derives multiple correction terms in the form of generalized inverse Mills ratios using bivariate normal integration. Monte Carlo simulation results show that the proposed MNP-TS estimator performs comparably or better in finite samples under asymmetric selection bias. The estimator also demonstrates computational stability as an alternative to full-information maximum likelihood, which may suffer from convergence problems due to the instability of high-dimensional likelihood functions. An empirical analysis of donation amounts using the 2024 Giving Korea data confirms the practical usefulness of the MNP-TS model.

Keywords Multinomial probit, sample selection bias, asymmetric bias, two-step estimator, Heckman model.

JEL Classification C31, C35, D64.

*Department of Economics, College of Political Science and Economics, Kyung Hee University, 26 Kyungheedaero, Dongdaemun-gu, Seoul, Republic of Korea 02447. E-mail: imin@khu.ac.kr.

비대칭적 표본선택 편의 교정을 위한 다항 프로빗 2단계 추정량

민인식*

Abstract 본 연구는 다항 선택(multinomial choice) 모형에서 발생하는 표본 선택 편의를 교정하기 위해 다항 프로빗 모형에 기반한 2단계 추정량(MNP-TS)을 제안한다. 기존의 다항 로짓 기반 교정법 Lee (1983)은 단일 확률 지수에 의존하는 교정항을 사용하여, 관측되지 않은 요인이 대안별로 비대칭적(asymmetric)인 편의를 유발할 때 추정 왜곡을 가져올 수 있다. 이러한 제약에 도전하기 위해 본 연구는 이변량 정규 적분을 활용하여 닫힌 형태(closed-form)의 일반화된 IMR (Generalized IMR)을 다중 교정항으로 도출하였다. 몬테카를로 시뮬레이션 결과, 제안된 MNP-TS 추정량은 비대칭적 선택 편의 환경에서 유사하거나 더 나은 유한표본 성과를 보였다. 또한 고차원 우도함수의 불안정성으로 수렴에 붕괴하는 완전정보 최우추정량(FIML)의 대안으로서 계산적 안정성을 입증하였다. ‘2024 기빙코리아’ 데이터를 활용한 기부 금액 결정요인 분석을 통해, MNP-TS 모형의 실증적 유용성을 확인하였다.

Keywords 다항 프로빗, 표본선택 편의, 비대칭적 편의, 2단계 추정량, 헤크만 모형.

JEL Classification C31, C35, D64.

*경희대학교 정경대학 경제학과 교수, 서울특별시 동대문구 경희대로 26 02447. E-mail: imin@khu.ac.kr.

1. 서론

종속변수가 모집단의 하위 집단에서만 관찰되고 이러한 선택의 무작위 과정이 아닌 경우에 표본선택 편의(sample selection bias)가 발생한다. Heckman (1976, 1979)은 이러한 과정을 무시하는 경우 OLS 추정량은 구조모형의 모수가 일치추정량(consistent estimator)이 되지 못함을 보여주었다. Heckman (1976)은 이항 선택(binary selection)인 모형에 기초하여 교정항(selection-bias correction term)을 포함하면 일치추정량을 얻을 수 있음을 증명하였다. Heckman의 2단계 추정에 따르면 선택(selection)과 결과(outcome) 방정식의 오차항에 대한 이변량 정규분포를 가정하고 1단계 추정에서 도출한 IMR(Inverse Mills Ratio)를 추가적인 설명변수로 포함하면 결과 방정식에서 조건부 잔차(conditional residuals)은 0이 되고 따라서 결과 방정식의 모수에 대해 일치추정량을 얻을 수 있다는 것이다.

그러나 1단계에 해당하는 선택 방정식에서 이항 종속변수로 가정하는 것은 실증분석 응용에서 제한적이다. 최근 표본선택 편의를 다루는 계량경제학 문헌은 이항선택 하에서 결과 변수가 다항 범주형(categorical outcome)인 경우의 편의 교정 Boussim (2025) 등 전통적 모형의 다각적인 확장을 시도하고 있다. 그러나 현실의 경제적 의사결정에서는 이와 반대로 선택 자체가 다항(multinomial)의 형태를 띠고, 그에 따른 연속형 결과 변수를 관측하게 되는 경우가 빈번하다.

가령, 노동경제학에서는 임금은 일자리를 가진 사람에게만 관찰되고 취업자의 직업 부문의 수는 $J \geq 3$ 일 수 있다. 어떤 부문을 선택하느냐와 임금과 상관관계가 있고, 이는 근로자의 관찰된/관찰되지 않은 특성에 의존한다 Willis and Rosen (1979); Dahl (2002). 또한, 기업의 수출 상품 가격은 ‘수출’을 선택한 기업에게만 관찰되며 그 가격은 어느 나라에 수출하느냐에 따라 달라질 수 있다 Helpman, Melitz and Rubinstein (2008). 사회복지의 기부 분야에서는 기부자에게만 기부금액이 관찰되며 기부금액은 기부 대상에 따라 달라질 수 있다. 가령 국내 기부(초록우산, 사회복지공동모금회)와 해외 기부(유니세프 등)로 구분할 수 있다. 관찰되지 않은 기부자 특성은 기부 여부와 기부금액을 동시에 결정한다고 이해할 수 있다.

Lee (1983)는 기존 Heckman 모형의 교정항을 다항 선택인 모형으로 체계적으로 확장한 최초 시도이다. 다항 로짓(Multinomial Logit: MNL) 1단계 모형과 표준정규 분포함수(CDF)의 역함수를 결합하여 Heckman의 IMR과 유사한 단변량 교정항을 도출하였다. Dubin and McFadden (1984)은 Lee (1983)의 단일 교정항이 보다 일반적인 교정 모형의 특수한 경우에 해당함을 보이고, MNL

오차항의 조건부 기댓값으로부터 직접 도출되는 $J-1$ 개의 교정항을 결과방정식에 포함하는 2단계 추정량을 제안하였다. 그러나 두 접근법 모두 1단계로 MNL을 사용하므로 IIA 가정을 전제하거나, 복잡한 다항 선택의 경계를 선택된 대안의 확률이라는 단일 지수로 축약하는 한계를 공유한다. Dahl (2002)은 joint normality 가정을 완화하여 선택교정 함수를 선택확률의 비모수적 함수로 근사하는 반모수적(semiparametric) 접근을 제시하였다. 분포 가정으로부터의 자유라는 장점이 있으나, 선택확률 추정에 여전히 MNL을 활용하는 경우 IIA 제약이 잔존하며, 비모수적 근사는 소표본에서 추정의 불안정성과 해석의 어려움을 수반한다.

이처럼 기존 접근법들은 공통적으로 대안 간 오차 의존성 구조를 제약하거나 단일 교정항으로 선택 편의를 축약하였다. 또한 실증 연구에서는 관찰되지 않은 이질성이 결과변수에 미치는 편의는 대안별로 비대칭적으로 나타날 수 있다. 기존의 교정항 방식은 이러한 비대칭적 편의를 개별적으로 식별하지 못하고 혼합하여 모수 추정에 왜곡을 가져올 가능성이 있다.

다항 프로빗(Multinomial Probit: MNP) 모형은 제약없는 공분산 구조를 갖는 다변량 정규분포를 가정함으로써 오차항 간 상관관계 구조를 포착할 수 있다 Hausman and Wise (1978); Daganzo (1979). Roodman (2007, 2011)이 개발한 Stata 명령어 `cmp`는 FIML(Full Information Maximum Likelihood)을 통해 결합 추정량을 제공하지만 고차원 우도함수의 불안정성으로 인해 수렴 실패의 부담이 있다.

명확한 실증적 동기와 추정 방법의 가용성에도 불구하고 1단계 MNP를 적용한 Heckman 교정에 대한 공식적인 계량경제학적 이론은 현재까지 상대적으로 다루어지지 않았다. 본 연구의 계량경제학적 기여를 다음과 같이 정리할 수 있다.

첫째, 다항로짓 모형은 오차항이 서로 독립적인 Type 1 Extreme Value 분포를 따른다고 가정한다. 따라서 대안 간 교차 공분산이 구조적으로 0으로 고정되어 IIA 가정이 성립하게 된다. 반면 다항 프로빗 모형은 오차항 벡터의 공분산 행렬 Σ (식 (3))를 비제약적(unrestricted) 형태로 추정하고 따라서 대안 사이의 non-zero 공분산이 허용된다. 이러한 가정은 체계적 의존성을 모형 내에서 직접적으로 포착하게 만든다. 따라서 IIA 가정이 위반되는 상황에서도 비제약 Σ 에 기반으로 도출된 2단계 추정량이 선택 편의를 흡수할 수 있게 된다. 이러한 과정은 IIA 가정에 의존하는 Lee (1983) 추정량에 비해 MNP-TS(Multinomial Two-Step)가 갖는 이론적 우위이다.

둘째, 선택 대안 갯수 $J = 3$ 이고 결합 정규성이 성립하는 MNP 모형에서 이변량정규 적분으로 축소 가능한 닫힌 형태(closed form)의 일반화된 IMR(Generalized

IMR: GIMR)을 도출한다. Heckman 모형의 단순한 교정항에 비해 MNP에서는 훨씬 복잡한 GIMR을 표현할 수 있다. 이러한 공식적인 도출은 기존 cmp에서 부재했던 계량경제학적 이론 기반을 제공한다.

셋째, 2단계 추정량에서 발생하는 생성된 설명변수(generated regressors) 문제를 논의하고 전체-파이프라인 부트스트랩을 통해 점근적으로 타당한 표준오차를 제공한다.

넷째, 비대칭적 선택편의가 존재하는 모형에서 MNP-TS 추정량의 유용성을 이론적 및 시뮬레이션을 통해 입증하고자 한다.

본 논문의 구성은 다음과 같다. II장에서는 1단계가 MNP인 표본선택 모형을 구체적으로 제안한다. III장에서는 다항로짓-2단계 추정량을 도출한다. IV장에서는 시뮬레이션 결과를 정리한다. V장에서는 ‘2024 기빙코리아’ 데이터를 활용한 실증분석 결과를 제시한다. 마지막 VI장에서 결론과 연구의 한계점을 서술한다.

2. 모형 설정

2.1. 선택 방정식과 결과 방정식

선택가능한 범주가 $J \geq 3$ 이고 각 선택 범주는 상호배반적이라고 가정한다. 개인 i 는 각 선택 j 에 대해 식 (1)에 따른 잠재효용을 갖게 된다.

$$s_{ij}^* = z_i \gamma_j + v_{ij}, \quad j = 0, 1, 2, \dots, J-1, \quad (1)$$

식 (1)에서 z_i 는 각 개인의 관찰된 특성이고 γ_j 는 선택가능한 대안(alternative-specific)에 대한 모수이다. 결과적으로 각 개인에 배정된 오차항 벡터 $v_i = (v_{i0}, v_{i1}, \dots, v_{iJ-1})$ 은 다변량 정규분포를 따르게 된다.

각 개인은 관찰된 선택은 선택가능한 대안 중에서 가장 높은 효용을 주는 대안이 된다.

$$s_i = \operatorname{argmax}_j \{s_{ij}^*\}.$$

MNP 모형을 식별하기 위해 정규화(normalization)를 적용한다. $j = 0$ (baseline) 선택할 때 효용을 빼준 값으로 설정한다. 식 (1)에 대한 차분 시스템은 다음과 같이 쓸 수 있다.

$$V_{ij}^* = z_i \delta_j + \bar{u}_{ij}, \quad j = 1, 2, \dots, J-1, \quad (2)$$

식 (2)에서 $V_{ij}^* = s_{ij}^* - s_{i0}^*$, $\delta_j = \gamma_j - \gamma_0$, $\bar{u}_{ij} = v_{ij} - v_{i0}$ 으로 정의한다. 식 (2)의 개인레벨 오차항 $\bar{u}_i$ 은 다음과 같은 분포를 따른다.

$$\bar{u}_i \sim \text{MVN}(0, \Sigma). \quad (3)$$

특정한 대안 j 를 선택하는 확률은 다음과 같이 쓸 수 있다.

$$P_{ij} = \Pr(s_i = j \mid z_i) = \Pr(V_{ij}^* > V_{ik}^* \text{ for all } k \neq j).$$

다항 프로빗 모형에서 선택확률은 $J-1$ 차원의 다변량 정규 확률로 쓸 수 있다. 이는 다항로짓의 확률 $P_{ij} = \frac{\exp(z_i \gamma_j)}{\sum_k \exp(z_i \gamma_k)}$ 와 차이가 있다. MNP는 식 (3)의 공분산 행렬 Σ 에 제약을 하지 않지만 MNL은 extreme value 오차항 분포에서 기인하는 대칭 구조를 암묵적으로 부과한다.

연속형 종속변수 y_i 는 대안 $J = 1, 2, \dots, J-1$ 을 선택한 개인에 대해서만 관찰된다. 관측된 종속변수를 갖는 개인에 대한 결과 방정식은 다음과 같다.

$$y_i = x_i \beta + u_i, \quad (4)$$

식 (4)에서 1단계 모형과 연결된 선택 편의가 없다면 $E(u_i \mid x_i) = 0$ 이 되고 분산은 $\text{var}(u_i) = \sigma_u^2 > 0$ 으로 가정할 수 있다. 식별 조건으로는 식 (1)과 식 (4)에서 $z \supset x$ 이 성립해야 한다. 즉 배제 제약(exclusion restriction)을 만족하는 변수가 z 에 적어도 하나 이상 포함되도록 설정한다.

2.2. 결합분포 및 선택 편의

선택편의를 구체적으로 정의하기 위해 결과 방정식 오차항 u_i 과 차분된 선택 방정식 오차항 $\bar{u}_i$ 가 다변량 정규분포를 따른다고 가정한다.

$$(u_i, \bar{u}_{i1}, \bar{u}_{i2}, \dots, \bar{u}_{iJ-1}) \sim \text{MVN}(0, \Omega), \quad (5)$$

식 (5)에서 Ω 는 $J \times J$ 공분산 행렬이고 $\Omega_{11} = \sigma_u^2$ 이고 우하단의 $(J-1) \times (J-1)$ 행렬은 Σ 가 된다. 좌하단의 $(J-1) \times 1$ 은 $\rho = (\rho_1, \dots, \rho_{J-1})$ 로 u_i 와 각 차분 오차항 $\bar{u}_j$ 와 공분산을 표현한다. 예상할 수 있듯이 선택 편의를 지배하는 구조 모수는 ρ 이다. 만약 $\rho = 0$ 이면 결과 방정식은 선택 과정과 독립적이고 따라서 결과 방정식인 식 (4)은 개별 선택 집단에 대한 OLS 추정량이 일치추정량이 된다. 만약 $\rho \neq 0$ 이면 선택 과정은 내생적이면 OLS 추정량은 일치추정량이 되지 못한다.

식 (5) 하에서 대안 m 을 선택한 경우, 결과방정식 오차항의 조건부 기댓값은 식 (6)과 같이 쓸 수 있다.

$$E(u_i | s_i = m, x_i, z_i) = \rho' \Sigma^{-1} E(\bar{u}_i | s_i = m, z_i), \quad (6)$$

식 (6)에서 $E(\bar{u}_i | s_i = m, z_i)$ 은 대안 m 이 선택되었다는 조건 하의 차분된 오차항들의 절단 평균(truncated mean)으로 해석할 수 있고, 이는 다변량 정규 분포함수로 표현할 수 있다.

$$\Lambda_i^m = (\Lambda_{i1}^m, \dots, \Lambda_{ij-1}^m) = E(\bar{u}_i | s_i = m, z_i), \quad (7)$$

식 (7)에 해당하는 Λ_i^m 이 대안 m 에 대한 GIMR이라고 해석할 수 있다. 식 (6)과 식 (7)를 식 (4)에 대입하면 다음과 같이 GIMR을 추가적인 설명변수로 포함하게 된다.

$$y_i = x_i \beta + \rho \Sigma^{-1} \Lambda_i^m + \varepsilon_i, \quad (8)$$

where $E(\varepsilon_i | s_i = m, x_i, z_i) = 0$.

식 (8)은 선택 편이가 교정된 결과 방정식의 구조적 형태(structural form)이다. 편의 교정항 Λ_i^m 은 전통적 Heckman 모형의 IMR을 다항 프로빗 환경에서 일반화한 값이다. 식 (8)을 하위 집단 $s_i = m$ 에 대해 OLS로 추정하면 β 에 대한 일치추정량을 얻을 수 있게 된다.

3. 2단계 추정량

FIML 추정량은 선택방정식과 결과방정식의 결합 로그우도함수를 동시에 최대화한다. 결합 정규성 가정 하에서 이론적으로 Cramer-Rao 하한을 달성하는 효율적인 추정으로 알려져 있다 Roodman (2011). 그러나 $J > 3$ 인 경우 다변량 정규 확률 계산에 수치적 시뮬레이션(GHK)이 요구되며 소표본에서는 수렴 불안정성이 발생할 수 있다. 본 연구에서는 FIML을 직접 구현하지 않고 Roodman (2011)의 Stata 사용자 명령어인 cmp를 활용하여 MNP-TS의 유한표본 성과와 비교한다¹.

본 연구에서 제안하는 2단계 추정량은 효율성을 희생하는 대신 계산적으로 더 단순한 대안을 제공하며 이는 전통적 Heckman 2단계 추정과 유사하다.

¹ Stata 명령문 `ssc install cmp, replace`로 설치할 수 있다.

3.1. 1단계: 다항 프로빗 모형

전체 N 개 관측치를 사용하여 Maximum Simulated Likelihood(최대 모의 우도)로 다항 프로빗 모형 식 (2)과 식 (3)를 추정한다. 그 결과 도출된 모수 추정치 $\hat{\delta}$ 와 $\hat{\Sigma}$ 를 얻게 된다. GHK 시뮬레이션 ($J > 3$) 또는 이변량 정규분포 확률 표 ($J = 3$)를 이용하여 아래 선택 확률을 계산할 수 있다.

$$\hat{\Pr}_{ij} = \Pr(s_i = j \mid z_i, \hat{\delta}, \hat{\Sigma}).$$

3.2. 2단계: GIMR 계산

다변량 정규분포의 truncated moments에 관한 Tallis (1961)과 Maddala (1983)의 논의를 다항 선택 문제로 확장하면 대안 m 을 선택한 개인 i 에 대해 k 번째 교정항은 다음과 같이 정의된다.

$$\hat{\Lambda}_{ik}^m = \frac{\partial \log P_i^m}{\partial (z_i \delta_k)} = \frac{\partial \hat{P}_i^m}{\partial (z_i \delta_k)} \times \frac{1}{\hat{P}_i^m}, \quad k = 1, 2, \dots, J-1, \quad (9)$$

식 (9)은 k 번째 선형함수 $z_i \delta_k$ 에 대한 로그 선택확률 $\log P_i^m$ 의 스코어(score)에 해당한다. 다항 프로빗의 경우 $\frac{\partial \hat{P}_i^m}{\partial (z_i \delta_k)}$ 은 다변량 정규 누적분포함수에 연쇄법칙을 적용하여 다시 쓸 수 있다 Geweke, Keane and Runkle (1994).

$$\frac{\partial \hat{P}_i^m}{\partial (z_i \delta_k)} = E [\phi(V_{ik}^*) \Phi_{J-2}(A_{ik}^m; B_{ik})], \quad (10)$$

식 (10)에서 $\phi(V_{ik}^*)$ 는 k 번째 임계값에서 평가된 단변량 표준 정규분포의 pdf이고 $\Phi_{J-2}(A_{ik}^m; B_{ik})$ 는 나머지 대안에 대한 $J-2$ 차원 표준정규분포 CDF이고 조건부 평균 A_{ik}^m 과 조건부 공분산 B_{ik}^m 을 이용하여 계산한다.

일반적으로 $J > 3$ 인 다항 선택모형에서는 다변량 적분을 계산하기 위해서는 GHK 시뮬레이션에 의존해야 한다. 그러나 본 연구에서 다루는 $J = 3$ 인 경우 대안 1($s = 1$)이 선택될 확률은 이변량 정규 적분으로 축소되고 따라서 수치적 근사 없이 닫힌 해(closed form)로 해석적 기울기 공식을 도출할 수 있다.

대안 1을 선택할 확률을 이변량 정규분포 식 $\Pr(s = 1) = \Phi_2(-a, -b; \rho_w)$ 여기서 $a = -z\delta_1/\Sigma_{w1}$, $b = (z\delta_2 - z\delta_1)/\Sigma_{w2}$ 이다². 이 확률에 대한 식 (13)의 기울기는 다음과 같이 도출된다.³.

² $\Phi_2()$ 는 이변량 누적분포함수를 의미한다.

³ $w_2 \equiv \bar{u}_{i1}$ (식 (3) 참고) 그리고 $w_2 \equiv (\bar{u}_{i1} - \bar{u}_{i2})$ 로 정의한다.

$$\frac{\partial \Pr(s=1)}{\partial z\delta_1} = \frac{\phi(a)}{\Sigma_{W_1}} \Phi\left(\frac{\rho_W a - b}{\sigma_\rho}\right) + \frac{\phi(b)}{\Sigma_{W_2}} \Phi\left(\frac{\rho_W b - a}{\sigma_\rho}\right), \quad (11)$$

$$\frac{\partial \Pr(s=1)}{\partial z\delta_2} = -\frac{\phi(b)}{\Sigma_{W_2}} \Phi\left(\frac{\rho_W b - a}{\sigma_\rho}\right). \quad (12)$$

따라서 결과 방정식에 포함되는 선택편의 교정항 $\hat{\Lambda}_{i1}^1 = \frac{\text{식 (11)}}{\Pr(s=1)}$ 그리고 $\hat{\Lambda}_{i2}^1 = \frac{\text{식 (12)}}{\Pr(s=1)}$ 이 된다.

3.3. 3단계: 확장된 결과 방정식 회귀모형

식 (11)과 식 (12)에서 구한 교정항을 포함한 후 선택된 하위 표본(가령 $(s_i = m)$)에 대한 OLS로 추정하는 모형은 다음과 같다. 아래 식에서 표기의 단순화를 위해 계수를 β 로 표현하였으나 이는 대안 m 을 선택한 하위 표본에 대해 독립적으로 추정되는 계수 벡터 β^m 을 의미한다.

$$y_i = x_i\beta + \alpha_1\hat{\Lambda}_{i1}^m + \alpha_2\hat{\Lambda}_{i2}^m + \cdots + \alpha_{J-1}\hat{\Lambda}_{iJ-1}^m + \varepsilon_i, \quad (13)$$

위 식에서 α_k 는 nuisance parameter이고 그 자체가 추론의 주된 관심 대상은 아니다⁴. 다변량 정규분포 하에서 score 벡터(식 (9))는 절단평균 벡터의 Σ^{-1} 선형변환이며, $J=3$ 인 경우 두 벡터가 동일한 공간을 span 하므로 여전히 식 (14)에 대한 OLS 추정량은 β 에 대한 일치추정량으로 유효하다.

식 (13)에서 다음과 같이 모든 교정항에 대한 결합 검정은 표본선택 편의에 대한 가설검정으로 해석할 수 있다.

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_{J-1} = 0,$$

위 가설을 기각한다면 1단계 선택 방정식을 무시하고 β 를 추정하는 것은 일치성에 문제가 될 수 있음을 시사한다.

MNP-TS는 식 (13)의 교정항이 1단계 MNP 추정치로부터 도출된 생성된 설명변수(generated regressors)이므로, 2단계 OLS 표준오차는 1단계 추정의 불확실성을 반영하지 못하여 일반적으로 과소추정된다(Pagan (1984)). 점근적 분산은 Murphy-Topel Murphy and Topel (1985) 형태의 샌드위치 공식으로 표현 가능하나, 교정항이 1단계 MNP 추정치의 비선형 함수이므로 해석적 유도

⁴식 (9)의 score 기반 교정항 $\hat{\Lambda}_{ik}^m$ 은 절단평균 벡터의 Σ^{-1} 의 선형변환이기 때문에 α_k 는 $\rho\Sigma^{-1}e_k$ 의 직접적 성분으로 해석되지 않는다.

가 쉽지 않다. 이에 본 연구에서는 1단계 추정의 불확실성을 전파하는 전체-파이프라인 부트스트랩(full-pipeline bootstrap)을 채택한다(Freedman and Peters (1984)). 부트스트랩은 각 반복(replication)에서 1단계 MNP 추정부터 2단계 OLS까지 전 과정을 재추정하므로, 분석적 샌드위치 분산 추정량을 명시적으로 도출하지 않고도 점근적으로 타당한 표준오차를 제공할 수 있다.

4. 시뮬레이션 결과

4.1. 시뮬레이션 설계

다음 네 가지 추정량의 유한 표본(finite sample) 특성을 평가하고자 한다. 1) 선택 편의를 무시한 순 OLS 2) 다항 로짓 1단계를 이용한 Lee (1983) 추정량 3) 본 연구에서 제안하는 2단계 MNP 교정 그리고 또 다른 일치추정량인 FIML 역시 벤치마크로 함께 보고한다. 모든 추정량은 선택 기제(selection mechanism)가 서로 다른 DGP(Data Generating Process) 하에서 평가된다.

논의를 단순화하기 위해 1단계 선택 가능한 대안은 $J = 3$ 인 모형을 설정한다. 개인 i 의 대안 1과 2에 대한 잠재 효용을 다음과 같이 정의한다. 대안 0이 기준 범주이다.

$$s_{1i}^* = \gamma_{10} + \gamma_{11}x_{1i} + \gamma_{12}x_{2i} + \delta_1 w_{1i} + v_{1i},$$

$$s_{2i}^* = \gamma_{20} + \gamma_{21}x_{1i} + \gamma_{22}x_{2i} + \delta_1 w_{2i} + v_{2i}.$$

관찰된 선택 $s_i = \operatorname{argmax} \{s_{0i}^*, s_{1i}^*, s_{2i}^*\}$ 으로 쓸 수 있다. 결과 방정식은 $s_i = 1$ 인 경우에 관측된다고 가정하자.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i \quad \text{if } s_i = 1, \quad (14)$$

식 (14)에서 조건부 기댓값 오차항 u_i 에 대한 조건부 기댓값은 다음과 같이 선택편의 교정항을 포함한다.

$$E(u_i | s_i = 1) = \alpha_1 E(v_{1i} | s_i = 1) + \alpha_2 E(v_{2i} | s_i = 1). \quad (15)$$

식 (15)의 $E(v_{1i} | s_i = 1), E(v_{2i} | s_i = 1)$ 은 저자가 작성한 Stata 코드 `mnpgimr.ado`를 통해서 계산되는 GIMR이 된다⁵.

⁵해당 코드는 교신저자에게 요청할 수 있다.

	DGP 1	DGP 2	DGP 3
모형 특징	오차항: MNL, IIA 성립	오차항: MNP, IIA 위반 심각	오차항: MNP, IIA 성립 (비대칭적 선택편의 가정)
Σ_{12}	0	0.80	0
MNP-TS	모형설정 오류	적절	적절
FIML	모형설정 오류	적절	적절
Lee (1983)	매우 적절	모형설정 오류	모형설정 오류

표 1: DGP에서 선택방정식의 오차항 분포. 이 표는 세 가지 DGP에서 선택 방정식의 오차항 분포와 각 추정량의 적합성을 비교하여 제시한다.

Table 1: ERROR DISTRIBUTIONS IN THE SELECTION EQUATION ACROSS DGPs. This table compares the error distributions in the selection equation and the suitability of each estimator across the three DGPs.

표 1에서는 세 가지 DGP에서 선택 방정식의 오차항 분포와 모수값을 정리하여 제시한다. 오차항 분포와 α 모수를 제외하고는 모두 동일하게 공유하며 이를 통해 오차항 구조와 비대칭적 선택편의가 추정량의 성능에 미치는 영향을 식별할 수 있도록 설계하였다.

특히 DGP1에서 선택모형 오차항 (v_{0i}, v_{1i}, v_{2i})는 Type-1 극단값 분포(Gumbel 분포)에서 상호 독립적으로 추출되고 따라서 MNL 선택 메커니즘을 유지한다. 이러한 DGP 하에서 MNP-TS와 FIML은 모형설정 오류 상황이 만들어진다. 이러한 분포 설정 오류에도 불구하고 본 연구에서 제안하는 MNP-TS 추정량의 강건성(robustness)을 테스트하는 역할을 할 수 있다. 표 2에서는 시뮬레이션에서 필요한 모수의 참값을 정리한다.

4.2. 추정량

베이스라인 추정량으로 선택편의가 없다고 가정하는 단순 OLS 추정량을 시도한다. 식 (14) 결과 방정식에서 $s_i = 1$ 하위 표본만 가지고 x_1 과 x_2 에 대해 회귀분석한다. 선택편의가 있는 모형에서는 $E(y_i | x_i, s_i = 1) \neq \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$ 이므로 OLS 추정량은 편의를 가지며 일치추정량이 되지 못한다.

Lee (1983) 추정량은 Heckman의 전통적 2단계 추정량을 다항 선택의 문제로 일반화한다. 1단계 선택방정식은 MNL 모형을 추정하여 예측확률 $\hat{p}_i = \hat{Pr}(s_i = 1 | x_i, w_{1i}, w_{2i})$ 을 얻는다. 2단계에서는 IMR을 다음과 같이 계산한다.

$$\hat{\lambda}_i^{Lee} = \phi(\Phi^{-1}(\hat{p}_i)) / \Phi(\Phi^{-1}(\hat{p}_i)),$$

모수	참값	설명
$\beta_0, \beta_1, \beta_2$	1.0, 1.0, 0.5	결과 방정식의 계수 (본 연구에서는 $\beta_1 = 1$ 을 추정)
α_1, α_2 (DGP1&2)	0.7, 0.5	GIMR 계수 (선택편의 교정항 계수)
α_1, α_2 (DGP3)	0.6, -0.7	
$\gamma_{10}, \gamma_{11}, \gamma_{12}, \delta_1$	0.5, 0.5, -0.3, 1.0	대안 1 효용함수의 계수
$\gamma_{20}, \gamma_{21}, \gamma_{22}, \delta_2$	0.3, -0.4, 0.5, 1.0	대안 2 효용함수의 계수
Σ_{12} (DGP1&3)	0	대안 1과 2의 오차항 상관관계 없음
Σ_{12} (DGP2)	0.8	대안 1과 2의 오차항 상관관계 매우 크다

표 2: 시뮬레이션 모수값. 이 표는 시뮬레이션에서 사용된 주요 모수의 참값과 그 의미를 제시한다.

Table 2: PARAMETER VALUES USED IN THE SIMULATION. This table presents the true values of the main parameters used in the simulation and their definitions.

3단계는 $s_i = 1$ 인 관측치를 이용하여 y 변수를 설명변수 $x_1, x_2, \hat{\lambda}^{Lee}$ 에 대해 OLS로 회귀분석 결과를 얻는다. 실제 분포가 MNL이라면 Lee 추정에서는 단일 교정항으로 축소시킨다. 그러나 실제 분포가 MNP인 경우에는 이변량 정규 구조를 정확히 활용하지 못하므로 MNP-TS에 비해 효율성이 떨어질 것으로 예상된다.

본 연구에서 제안하는 MNP-TS의 1단계는 설명변수 x_1, x_2, w_1, w_2 를 포함하여 최우추정법으로 MNP 모형을 추정한다. 추정된 오차항 상관계수 $\hat{\rho}_{12}$ 를 얻게 된다. 2단계에서는 저자가 코딩한 `mnp_gimr` 명령어를 이용하여 GIMR을 계산한다 (식 (9) 참고). 3단계에서는 $s = 1$ 인 관측치를 이용하여 종속변수 y 를 설명변수 $x_1, x_2, \hat{\Lambda}_1^{s=1}, \hat{\Lambda}_2^{s=1}$ 에 대해 OLS로 회귀분석을 한다. 여기서 교정항의 계수는 식 (15)의 α_1 과 α_2 에 해당한다.

FIML 추정량은 $s = 1$ 인 관측치의 연속형 결과 방정식과 전체 관측치의 다항 프로빗 선택방정식을 결합한 완전 데이터(complete data)의 우도를 최대화한다. 이는 Stata에서 Roodman (2011)이 개발한 `cmp` 명령어를 사용하여 구현할 수 있다. 본 시뮬레이션에서 FIML은 효율성 벤치마크의 역할을 한다. 실제 분포가 MNP로 설정된 경우 FIML은 점근적으로 Cramer-Rao bound를 달성한다. MNP-TS는 일치성을 갖지만 2단계 추정이라는 본질적인 제약으로 인해 FIML에 비해 효율성이 떨어진다. MNP-TS와 FIML을 비교함으로써 효율성 손실을 정량화할 수 있다.

4.3. 시뮬레이션 설정과 결과

본 연구에서 시뮬레이션은 DGP1, DGP2 그리고 DGP3에 대해 각각 표본 크기 300, 500, 1,000개로 총 9번의 시뮬레이션 결과를 얻는다. 앞선 V장 2절에서 제시한 4개 추정량 각각에 대해 시뮬레이션 결과를 정리한다. 따라서 각 시뮬레이션 결과는 ‘추정량-DGP-표본크기’ 조합에 대해 식 (17)의 참값($\beta_1 = 1$)과 비교한다.

평가 기준은 일반적으로 시뮬레이션 분석에서 채택하는 편의(Bias)와 RMSE를 계산한다.

$$Bias = \frac{1}{R} \sum_r (\hat{\beta}_{1r} - 1.0), \quad (16)$$

$$RMSE = \sqrt{\frac{1}{R} \sum_r (\hat{\beta}_{1r} - 1.0)^2}, \quad (17)$$

where R은 반복횟수(the number of repetitions)를 의미한다.

추가적으로 FIML 경우는 수렴 비율(convergence rate)을 핵심 평가지표로 보고한다. FIML은 선택과 결과 방정식의 모두 모수, 그리고 오차항의 공분산 행렬을 고차원의 모수 공간에서 결합 우도함수를 동시에 극대화해야 한다. 이러한 과정에서 우도함수는 수치적 최적화 알고리즘 수렴에 실패할 가능성이 구조적으로 존재한다. 비록 FIML이 MNP 가정 하에서 점근적으로 가장 효율적인 추정량일지라도 현실의 실증분석 데이터 환경에서 수렴을 달성하지 못한다면 그 유용성은 제한되기 때문이다.

표 3에서는 시뮬레이션 결과를 각 DGP별로 정리한 결과를 제시한다. 먼저 선택편의를 고려하지 않은 OLS 추정량은 $N = 300$ 에서 $N = 1000$ 으로 증가하더라도 편의(bias)가 줄어들지 않는다. 이는 선택 오차항과 결과 방정식의 오차항 간 상관관계를 통제하지 않으면 추론의 오류가 발생할 수 있음을 재확인한다.

MNL이 참 모형인 DGP1에서 MNP-TS 추정량은 분포 설정오류임에도 불구하고 Lee 추정량과 거의 대등한 수준의 편의 교정 능력을 보여주며 높은 강건성을 입증한다. 대안 간 상관관계가 존재하는 DGP2에서도 MNP-TS는 표본 크기가 증가함에 따라 편의가 0에 가깝게 수렴($N=1000$, -0.003) 시키며 일치성을 확인할 수 있다. 다만 현재의 모수 설계 하에서 Lee 추정량과의 유한 표본 성과 차이가 충분히 발현되지 않은 것은 본 연구의 한계로 보인다.

본 연구의 핵심인 비대칭적 선택편의를 가정한 DGP 3의 결과는 다중 교

정향을 사용하는 MNP-TS는 이론적으로 올바르게 설정된 추정량으로서 편의 교정에 있어 Lee 추정량과 유사하거나 더 나은 유한표본 성과를 보인다. 편의의 원천이 하나의 대안으로 치우친 상황에서 단일 교정항에 의존하는 Lee 추정량은 양(+)과 음(-)의 오차 구조를 명확히 분리하지 못해 편의가 상대적으로 크다. 더 효율적인 추정량으로 알려진 FIML 역시 $N=300$ 에서 $\text{Bias}=-0.046$ 의 MNP-TS의 3배 이상이다. 비대칭적 공분산 구조 하에서 결합 우도함수가 평탄해져 수치적 최적화가 왜곡될 수 있음을 시사한다.

MNP-TS는 100%의 계산적 안정성(수렴 비율)을 보장하고 동시에 두 개의 교정항 $\hat{\Lambda}_1^{s=1}, \hat{\Lambda}_2^{s=1}$ 을 통해 비대칭적 편의를 개별적으로 통제하고 모든 표본 크기에서 가장 낮은 편의와 RMSE를 달성하였다. 결론적으로 MNP-TS는 FIML의 수렴 한계와 기존 Lee 모형의 구조적 한계에 도전하며 안정적인 2단계 추정량임을 확인할 수 있다.

5. 실증분석: 기부영역과 기부금액

본 장에서는 다항 프로빗 2단계(MNP-TS) 추정량의 실제 활용 가능성을 보여주기 위해 한국의 기부 데이터를 이용한 실증분석을 진행한다. 구체적으로 개인의 기부 대상 선택(국내기부 vs. 해외기부)과 기부금액 결정 모형을 추정하며 특정 기부 대상을 선택하게 만드는 관측되지 않은 요인이 결과변수(기부금액)에 미치는 비대칭적 표본선택 편의 문제와 이를 통제하기 위한 다중 교정항 역할에 초점을 맞춘다.

실증분석을 위한 데이터는 아름다운재단 기부문화연구소가 주관하고 한국갤럽이 수행한 『2024년 기빙코리아(Giving Korea)』 조사 데이터를 사용한다⁶. 기빙코리아 조사는 2001년부터 격년으로 실시되어 온 한국의 대표적인 기부 및 자원봉사 실태조사이다. 개인의 기부 행태, 인식, 그리고 인구·사회학적 특성에 대한 광범위한 정보를 제공한다. 실증분석을 위해 해당 조사의 raw data 중 개인기부 문항을 추출하여 사용하였다.

1단계 선택 방정식은 개인이 기부 대상을 선택하는 의사결정 과정으로, 종속변수 s_i 는 표 4와 같이 세 가지 상호배타적인 대안을 갖도록 설계하였다.

2단계 결과 방정식에서 종속변수 y_i 는 기부 경험이 있는 사람의 금전적 금액(물품 기부 제외)의 로그 값으로 설정된다. $s_i = 1, s_i = 2$ 를 선택한 표본에 대해서만 관찰된다. 식 (18)와 식 (19)에서는 모형 설정의 구체적인 형태를 보여준다. X_i 는 기부 여부와 기부 금액에 모두 영향을 미치는 개인 특성(성별, 연령, 소득 등) 설명변수를 포함한다.

⁶https://research.beautifulfund.org/archive/?cat=922&slug=giving_korea_rawdata

패널 A: DGP1 (MNL 분포 & IIA 가정 성립)						
	$N = 300$		$N = 500$		$N = 1000$	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
OLS	-0.106	0.180	-0.106	0.154	-0.113	0.137
Lee	0.030	0.167	0.029	0.132	0.017	0.089
MNP-TS	0.032	0.173	0.029	0.134	0.015	0.090
FIML	0.016	0.149	0.030	0.138	0.009	0.093
FIML conv.rate	0.55		0.62		0.77	
패널 B: DGP2 (MNP 분포 & IIA 가정 위반 심각)						
	$N = 300$		$N = 500$		$N = 1000$	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
OLS	-0.065	0.145	-0.062	0.124	-0.070	0.102
Lee	0.017	0.142	0.014	0.114	0.005	0.081
MNP-TS	0.008	0.143	0.003	0.117	-0.003	0.081
FIML	0.004	0.143	0.002	0.113	-0.0005	0.081
FIML conv.rate	0.45		0.60		0.81	
패널 C: DGP3 (MNP 분포 & 비대칭적 선택 편의)						
	$N = 300$		$N = 500$		$N = 1000$	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
OLS	-0.164	0.202	-0.165	0.189	-0.148	0.160
Lee	-0.021	0.130	-0.017	0.109	-0.003	0.070
MNP-TS	-0.015	0.130	-0.011	0.108	-0.002	0.070
FIML	-0.046	0.134	-0.036	0.116	-0.012	0.074
FIML conv.rate	0.67		0.79		0.93	

표 3: 시뮬레이션 결과 정리. 이 표는 세 가지 DGP와 표본 크기별로 각 추정량의 Bias, RMSE 및 FIML 수렴 비율을 제시한다. 주) 반복횟수 $R = 300$ 회를 진행.

Table 3: SUMMARY OF SIMULATION RESULTS. This table presents the bias, RMSE, and FIML convergence rate for each estimator across the three DGPs and sample sizes.

Note: The number of repetitions is $R = 300$.

$$\text{선택방정식: } s_{ji}^* = \gamma_{j0} + \gamma_j X_i + \delta_j W_i + v_{ji}, \quad (18)$$

$$\text{결과방정식: } y_{ji} = \beta_{j0} + \beta_j X_i + u_{ji}, \quad (19)$$

s_i	정의
0	기부 미참여: 2023년 한 해 동안 금전적 기부(종교적 헌금 제외) 없는 집단
1	기부 경험이 있으며 ‘국내 관련 분야’로 한정 (국내 사회복지, 교육, 병원, 문화예술, NGO 등)
2	기부 경험이 있으며 ‘해외 관련 분야’로 한정 (외국의 의료, 자연재해 지원)

표 4: 선택방정식의 선택 대안: 정의. 이 표는 선택 방정식에서 사용된 세 가지 선택 대안의 정의를 제시한다. 주) 국내와 해외에 동시에 기부한 사람은 분석대상에서 제외함.

Table 4: DEFINITIONS OF ALTERNATIVES IN THE SELECTION EQUATION. This table presents the definitions of the three alternatives used in the selection equation. Note: Individuals who donated to both domestic and overseas causes are excluded from the analysis.

where $j = 1, 2$.

위 실증분석 모형에서 MNP-TS 추정량의 적용이 갖는 계량경제학적 기여는 다음과 같다. 전형적인 횡단면 설문조사인 기빙코리아 데이터는 개인특성 변수(case-specific variable)으로 구성되어 있어 대안 간 오차항 공분산을 식별하기에는 정보의 한계가 존재한다. 그러나 앞선 시뮬레이션(DGP3)에서 확인하였듯이 관찰되지 않은 특성은 결과 변수에 미치는 영향이 대안별로 비대칭적인 경우 기존 Lee의 단일 확률지수(single probability index) 교정법은 편의를 유발할 수 있다. 현실의 기부 의사결정에서 ‘국내 기부’를 유인하는 관습적·지역적 동기와 ‘해외 기부’를 유발하는 글로벌 이슈에 대한 민감도는 실제 지출하는 기부금액의 크기에 비대칭적 파급효과를 미칠 가능성이 있다. 따라서 본 장에서는 두 개의 독립적인 GIMR을 동시에 포함하여 대안별 선택 편의를 분리·통제하는 MNP-TS 추정량이 현실 데이터에서 보여주는 실제적 유용성을 확인하고자 한다.

모형 식별 전략으로는 식 (18)의 W 변수를 포함하도록 한다. 즉, 결과방정식에는 포함되지 않으면서 기부 참여 및 대상 선택에만 영향을 미칠 수 있는 배제 제약 변수가 필요하다. 기빙코리아 설문지 문항을 바탕으로 살펴보면 기부 금액 크기보다 ‘기부 참여 자체’에 더 직접적인 영향을 미칠 것으로 예상되는 변수는 ‘기부 관련 정보 습득 채널’이다. 가령 ‘TV 또는 라디오 방송프로그램이나 광고’를 접촉한 경험 유무, ‘거리모금을 통한 기부 정보 획득, 포털사이트 기부플랫폼 활용 여부(해피빈 등)’를 예로 들 수 있다. 노연희·전현경 (2017). 특히 한국 기부 시장에서 ‘거리 모금’은 유니세프, 세이브더칠드런 등 대형 국제구호 NGO가 압도적으로 주도하고 있는 모금 방식이다 강철희·구지윤·박소현 (2011). 길거리에서 기부 정보를 접한 개인은 해외기부 $s = 2$ 를 선택할 확률이 높아질 수 있다.

방정식	구분	변수
선택방정식	종속변수 s	기부미참여=0, 국내기부만=1, 해외기부만=2
	X	성별, 연령, 가구소득, 거주지역, 종교 유무, 혼인상태, 근로 여부, 교육수준
	W_1	포털사이트 기부 플랫폼 여부 또는 TV, 라디오 방송프로그램 여부 (국내기부 식별)
결과방정식	W_2	거리모금 여부 (해외기부 식별)
	종속변수 y	기부금액의 로그값 (기부를 한 사람만 관찰됨)
	X	성별, 연령, 가구소득, 거주지역, 종교 유무, 혼인상태, 근로 여부, 교육수준

표 5: 실증분석 모형에 사용된 변수. 이 표는 선택방정식과 결과방정식에서 사용된 종속변수 및 설명변수를 제시한다.

Table 5: VARIABLES USED IN THE EMPIRICAL MODEL. This table presents the dependent and explanatory variables used in the selection and outcome equations.

표 5에서는 실증분석에 사용된 종속변수와 설명변수를 요약하여 정리한다. 선택방정식 중 국내기부를 식별하기 위해 변수 W_1 은 [포털사이트 기부 플랫폼 또는 TV, 라디오 방송프로그램]을 더미변수로 포함한다. 해외기부를 식별하기 위한 변수 W_2 는 [거리모금] 경험 여부로 설정한다.

표 6은 세 가지 선택 대안별 주요 변수의 기초통계량을 제시한다. 전체 표본 중 기부 미참여 비율이 53.3%로 과반을 차지하며, 국내기부만 선택한 비율은 41.6%, 해외기부만 선택한 비율은 5.1%로 나타난다. 기부금액은 국내기부 집단과 해외기부 집단 간에 차이가 관찰되며, 이는 두 집단의 기부 행태가 상이할 수 있음을 시사한다.

선택 편의에 대한 교정이 없는 OLS 추정과 본 연구에서 제안한 MNP-TS 추정량 결과는 표 7에서 정리하여 제시한다. 분석 대상은 기부 대상에 따라 국내기부 $s = 1$ 과 해외기부 $s = 2$ 표본으로 나누어 수행하였다. 저자가 코딩한 Stata 명령어 `mnp_heckman.ado`을 이용한 결과를 정리한다⁷. MNP-TS 추정량의 표준오차는 부트스트랩(bootstrap) 방법으로 산출하였다.

국내기부자 표본의 분석 결과 단순 OLS와 MNP-TS 모형 간 추정계수 방향성은 전반적으로 일치하나 통계적 유의성에서 차이가 있다. 단순 OLS 모형에서는 연령(age), 직업 유무(job), 혼인상태(married), 종교(religion), 소득(log income)이 기부금액에 유의하게 영향을 미친다. 그러나 MNP-TS에서는 교육수준이 1% 유의수준에서 기부금액을 유의하게 증가시킨다. 반면 연령의 유

⁷ 해당 코딩파일은 교신저자에게 요청할 수 있다.

변수	기부 미참여 (53.3%)		국내기부만 (41.6%)		해외기부만 (5.1%)	
	평균	표준편차	평균	표준편차	평균	표준편차
기부금액(만 원)	—	—	23.1	54.1	12.5	14.7
성별(남자 비율)	0.48	0.50	0.54	0.49	0.42	0.49
연령	47.7	15.7	48.3	15.1	47.6	15.3
가구소득(백만 원)	4.3	4.9	4.9	5.0	4.4	3.0

표 6: 주요 변수의 기초통계량. 이 표는 선택 대안별 주요 변수의 평균과 표준편차를 제시한다.

Table 6: DESCRIPTIVE STATISTICS OF MAIN VARIABLES. This table presents the means and standard deviations of the main variables by choice alternative.

의성은 상실되었고 종교의 유의수준은 하락하였다. 이러한 결과는 단순 OLS 모형의 오차항에 포함되어 있던 교란 요인을 통제하여 기부금액 결정에 대한 인과적 영향을 포착하고 있음을 보여준다. MNP-TS 모형에 포함된 GIMR 교정항은 모두 통계적으로 유의하지 않다. 이는 국내 기부 표본 내에서는 기부참여를 결정하는 관찰되지 않은 특성이 기부금액에 미치는 내생성이 통계적으로 심각한 수준이 아니라고 해석할 수 있다

해외기부자 표본에서는 선택 편의 교정 여부에 따라 변화가 발생한다. 단순 OLS 모형에서는 종교(religion)와 연령(age) 변수가 기부금액에 유의한 양(+)의 효과를 미친다. 그러나 MNP-TS 모형에서는 해당 변수의 유의성이 모두 사라진다. 다만, 이러한 통계적 상실을 표본선택 편의 교정 결과로만 판단하기에는 무리가 있다. 해외 기부자 표본 수(N=96)가 소표본이라는 환경에 기인했을 가능성이 크다 (추가적인 자유도 손실).

6. 결론과 연구의 한계

본 연구는 다항 선택 구조에서 발생하는 선택 편의를 교정하기 위해 기존 단일 교정항 방법론 Lee (1983)의 구조적 한계에 도전하는 다항 프로빗 기반의 2단계 추정량(MNP-TS)을 제안하고 그 유한 표본 특성을 논의하였다. 실증 데이터에서 관찰될 수 있는 비대칭적 선택 편의 상황에서 단일 확률 지수에 의존하는 기존 모형은 양(+)과 음(-)의 교란 요인을 적절하게 분리하지 못한 문제를 발견하였다. 반면 본 연구에서 제안한 MNP-TS 추정량은 $J-1$ 개 다중 교정항을 결과 방정식에 포함하여 비대칭적 편의를 통제할 수 있다.

몬테카를로 시뮬레이션 결과, MNP-TS 추정량은 고차원 우도함수의 수치

변수	$s = 1$ (국내기부만)		$s = 2$ (해외기부만)	
	OLS	MNP-TS	OLS	MNP-TS
male (남자)	0.002 (0.108)	-0.067 (0.159)	0.335 (0.317)	0.233 (0.347)
reg3=2 (기타 광역시)	-0.142 (0.128)	-0.101 (0.165)	-0.562 (0.364)	-0.531 (0.538)
reg3=3 (기타 시도)	-0.354 (0.133)***	-0.305 (0.173)**	-0.399 (0.388)	-0.371 (0.304)
edu3=2 (고졸)	2.257 (1.491)	2.269 (0.408)***	0.212 (0.987)	-0.045 (1.190)
edu3=3 (대졸이상)	2.311 (1.487)	2.327 (0.400)***	0.550 (0.919)	0.242 (1.247)
job=1 (일자리 있음)	0.421 (0.124)***	0.450 (0.187)***	0.239 (0.361)	0.067 (0.673)
married=1 (배우자 있음)	0.553 (0.132)***	0.519 (0.210)***	0.388 (0.362)	0.316 (0.548)
religion=1 (종교 있음)	0.390 (0.110)***	0.485 (0.201)*	0.698 (0.324)**	0.456 (1.014)
age	0.008 (0.004)**	0.008 (0.005)	0.021 (0.012)*	0.021 (0.016)
log(income)	0.042 (0.014)***	0.042 (0.016)***	0.042 (0.042)	0.031 (0.061)
$\hat{\Lambda}_1^s=1$	-	-0.423 (1.105)	-	-
$\hat{\Lambda}_2^s=1$	-	3.096 (2.943)	-	-
$\hat{\Lambda}_1^s=2$	-	-	-	1.297 (2.125)
$\hat{\Lambda}_2^s=2$	-	-	-	2.110 (1.619)
obs	783	783	96	96

표 7: 기부금액 모형 추정결과(결과 방정식). 이 표는 국내기부 및 해외기부 표본에 대한 OLS와 MNP-TS 추정결과를 제시한다. 주1) reg3=1(수도권), edu3 =1(중졸 이하)가 기준 범주임. 주2) 괄호 안의 값은 추정계수의 표준오차이고, 추정된 교정항의 불확실성을 고려하기 위해 부트스트래핑(bootstrapping) 100회를 반복하여 얻었다. 주3) ***, **, *은 각각 1%, 5%, 10% 수준에서 유의함을 의미한다.

Table 7: ESTIMATION RESULTS FOR THE DONATION AMOUNT MODEL (OUTCOME EQUATION). This table presents the OLS and MNP-TS estimation results for the domestic-donation and overseas-donation samples. Notes: reg3=1 (Seoul metropolitan area) and edu3=1 (middle school or below) are the reference categories. Standard errors are reported in parentheses and are obtained from 100 bootstrap replications to account for the uncertainty of the estimated correction terms. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

적 불안정성으로 인해 수렴에 실패할 수 있는 FIML의 실용적 한계를 극복할 수 있었다. 아울러 비대칭적 선택 편의 가정(DGP3)에서는 Lee 추정량보다 유사하거나 더 나은 유한표본 성과를 보였다. IIA 가정이 위반된 환경에서도 일치추정량으로서 이론적 타당성을 유지한다.『2024 기빙 코리아』 기부 데이터를 활용한 기부금액 모형 추정을 통해 다항 선택과정을 통제한 MNP-TS 모형이 교육수준 변수의 계수 추정에 영향을 미침을 확인하였다.

다만 본 연구가 제안한 MNP-TS 방법론은 다음과 같은 구조적 한계점을 지닌다. 수치적 불안정성을 줄이기 위해 도출한 GIMR의 해석적(analytical) 기울기 공식은 이변량 정규 적분에 기반한다. 따라서 $J = 3$ 인 경우에만 적용가능하다. 대안이 4개 이상인 $J > 3$ 인 다항선택 환경으로 2단계 추정량을 확장하기 위해서는 다변량 정규 확률의 편미분 근사를 위해 고도화된 GHK 시뮬레이션 알고리즘을 제안하는 후속 연구가 요구된다.

참고문헌

- 강철희 · 구지윤 · 박소현 (2011). “기부지역 선택행동에 대한 영향요인 탐색: 국내기부와 해외기부의 비교를 중심으로,” *사회복지정책* 38, 221–253.
- (Translated in English) Kang, C., J. Goo, and S. Park (2011). “An exploratory study on factors influencing donation-area choice: Focusing on a comparison of domestic and overseas donations,” *Social Welfare Policy* 38, 221–253.
- 노연희 · 전현경 (2017). “온라인 모금캠페인의 내용 및 성과의 탐색적 분석: 같이가치 with Kakao의 모금캠페인을 중심으로,” *소비자학연구* 28, 171–197.
- (Translated in English) Noh, Y. and H. Jeon (2017). “An exploratory analysis of the content and performance of online fundraising campaigns: Focusing on the Together with Kakao fundraising campaigns,” *Journal of Consumer Studies* 28, 171–197.
- Amemiya, T. (1985). *Advanced Econometrics*, Harvard University Press.
- Boussim, O. (2025). “Correcting sample selection bias with categorical outcomes,” arXiv preprint arXiv:2510.05551.
- Daganzo, C. (1979). *Multinomial Probit: The Theory and Its Application to Demand Forecasting*, Academic Press.
- Dahl, G.B. (2002). “Mobility and the return to education: Testing a Roy model with multiple markets,” *Econometrica* 70, 2367–2420.
- Dubin, J.A. and D.L. McFadden (1984). “An econometric analysis of residential electric appliance holdings and consumption,” *Econometrica* 52, 345–362.
- Freedman, D.A. and S.C. Peters (1984). “Bootstrapping a regression equation: Some empirical results,” *Journal of the American Statistical Association* 79, 97–106.
- Geweke, J. (1991). “Efficient simulation from the multivariate normal and Student-*t* distributions subject to linear constraints and the evaluation of constraint probabilities,” *Computer Science and Statistics, Proceedings of the 23rd Symposium on the Interface*, 571–578.

- Geweke, J., M. Keane and D. Runkle (1994). "Alternative computational approaches to inference in the multinomial probit model," *The Review of Economics and Statistics* 76, 571–578.
- Hausman, J.A. and D.A. Wise (1978). "A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous preferences," *Econometrica* 46, 403–426.
- Heckman, J.J. (1976). "The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models," *Annals of Economic and Social Measurement* 5, 475–492.
- Heckman, J.J. (1979). "Sample selection bias as a specification error," *Econometrica* 47, 153–161.
- Helpman, E., M.J. Melitz and Y. Rubinstein (2008). "Estimating trade flows: Trading partners and trading volumes," *Quarterly Journal of Economics* 123, 441–487.
- Lee, L.F. (1983). "Generalized econometric models with selectivity," *Econometrica* 51, 507–512.
- Maddala, G.S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge University Press.
- Murphy, K.M. and R.H. Topel (1985). "Estimation and inference in two-step econometric models," *Journal of Business and Economic Statistics* 3, 370–379.
- Newey, W.K. and D. McFadden (1994). "Large sample estimation and hypothesis testing," in R.F. Engle and D.L. McFadden (eds), *Handbook of Econometrics*, Vol. 4, Elsevier, 2111–2245.
- Pagan, A. (1984). "Econometric issues in the analysis of regressions with generated regressors," *International Economic Review* 25, 221–247.
- Roodman, D. (2007). "cmp: Stata module to implement conditional (recursive) mixed process estimator," *Statistical Software Components S456882*, Department of Economics, Boston College.

- Roodman, D. (2011). “Fitting fully observed recursive mixed-process models with cmp,” *The Stata Journal* 11, 159–206.
- Tallis, G.M. (1961). “The moment generating function of the truncated multi-normal distribution,” *Journal of the Royal Statistical Society: Series B (Methodological)* 23, 223–229.
- Willis, R.J. and S. Rosen (1979). “Education and self-selection,” *Journal of Political Economy* 87, S7–S36.
- Wooldridge, J.M. (2010). *Econometric Analysis of Cross-Section and Panel Data*, MIT Press.