Journal of the Korean Data & Information Science Society
2017, **28**(5), 1077–1085

http://dx.doi.org/10.7465/jkdi.2017.28.5.1077
한국데이터정보과학회지

표본 선택 모형을 이용한 국내 여성 임금 데이터 분석[†]

정미량[1] · 김미정[2]

[12]이화여자대학교 통계학과

접수 2017년 8월 17일, 수정 2017년 9월 13일, 게재확정 2017년 9월 14일

요 약

본 연구에서는 한국노동연구원의 「2015년 한국노동패널조사 (KLIPS)」자료를 활용하여 국내 여성의 임금 결정요인을 분석하기로 한다. 일반적으로 임금 자료는 랜덤 추출이 불가능하기 때문에 분석하기가 쉽지 않다. 표본 선택 편의 (sampling bias)가 있는 자료를 분석하는 방법으로 Heckman 표본 선택 모형이 가장 널리 알려져 있다. Heckman은 크게 두 가지 모형을 제안했는데, 그 중 하나는 최대 우도 방법을 이용하는 것이고, 다른 하나는 2단계 표본 선택 모형이다. 이 중 Heckman 2단계 표본 선택 모형은 주된 결과 모형 (outcome model)과 경제 활동 여부를 결정짓는 선택 모형 (selection model)을 포함한 모형으로써, 이 모형이 최대 우도 방법을 이용한 모형에 비해 이변수 오차의 정규분포 가정에 덜 민감하다고 알려져 있다. 그럼에도 불구하고 이변수 오차에 대한 정규 분포 가정은 꽤 강한 가정이라고 볼 수 있는데, 최근에 이 모형의 단점을 보완하는 모형으로 Marchenko와 Genton (2012)의 Heckman 표본 선택 t 모형이 제시되었다. Heckman 2단계 모형과 Heckman 표본 선택 t 모형을 이용하여 국내 여성의 임금 결정 요인을 분석하고 비교하도록 한다.

주요용어: 결측 자료, 표본 선택 편의, 헤크만 모형, 헤크만 표본 선택 t 모형.

1. 서론

모집단의 특성을 파악하기 위해 모집단을 대표할 수 있는 표본을 추출하는 것은 데이터 분석 및 예측 측면에서 중요한 일이다. 그러나 때로는 대표성이 있는 표본을 랜덤하게 추출하는 것이 어려운 경우가 빈번하게 발생한다. 대표적인 예로, 1900년대 초 미국에서 시행된 대통령 선거 전 전화 설문 전화를 들 수 있다. 그 당시 전화를 소유한 사람은 주로 부유층이었기 때문에, 이 때 시행된 전화 설문 조사는 전 국민의 의견을 대표하기 보다는 부유층의 의견을 대표했다고 볼 수 있다. 임금 데이터에 수집 시에도 랜덤한 표본을 구하기는 어렵다. 사람들이 일을 하지 않은 이유가 랜덤하게 발생하지 않기 때문이다. 따라서 임금을 결정짓는 원인을 분석하는 것은 어려운 일이었으나, Heckman은 표본 편의 (selection bias)가 존재하는 임금을 분석할 수 있는 방법인 Heckman two-stage 방법을 고안하였다. Heckman은 이러한 공로를 인정받아 2000년에 노벨 경제학상을 받았다. Park과 Cho (2016)은 Heckman이 제시한 모형을 이용하여 표본 선택 편의를 반영한 임금 결정 요인을 분석하였다. 표본 편의가 존재하는 자료를 분석하는 방법으로 Heckman이 시도한 방법은 획기적이었음이 분명하지만, 특정 가정이 강하다는 단점이 있고, 이러한 단점을 보완하여 Marchenko와 Genton (2012) 에 의해 Heckman 표본 선택 t 모형이

[†] 이 논문은 2017년도 한국연구재단 연구비 (과제 번호: NRF-2017R1C1B5015186)에 의하여 수행되었음.
[1] (03760) 서울특별시 서대문구 이화여대길 52 이화여자대학교 통계학과 대학원.
[2] 교신저자: (03760) 서울특별시 서대문구 이화여대길 52 이화여자대학교 통계학과 조교수.
E-mail: m.kim@ewha.ac.kr

고안하였다. 이 논문에서는 이러한 두 가지 방법을 이용하여「2015년 한국노동패널조사」데이터를 통해 2015년 기혼여성의 경제활동인구 여부와 임금결정요인에 대해 분석해 보기로 한다. 논문의 구성은 다음과 같다. 제 2절에서 본 논문에서 분석에 사용된 데이터에 대한 소개와 기술통계에 대한 내용과 제 3절에서 표본선택모형에 대하여 간단히 설명한다. 제 4절은 제 3절에서 소개한 표본선택모형에 데이터를 적용시킨 결과를 제시하고, 마지막 제 5절은 결론을 제시한다.

2. 자료 소개

통계청에서 발표한「경제활동인구조사」에 따르면, 여성의 경제활동인구는 40% 초반으로 해가 지날수록 경제활동 비율이 점차 증가하는 것을 알 수 있다. Table 2.1은 전체 경제활동인구 중 여성의 경제활동인구 비율이다. 경제활동인구조사는 가구원별 일자리에 대한 내용, 구직에 대한 내용 등으로 구성되어 있다. 본 논문에서 활용할「한국노동패널조사」는 가구의 소득, 소비, 자산 등 가구 또는 개인에 대하여 세분화된 내용을 포함하고 있다는 점과 종단적인 측면에서 분석을 할 수 있다는 점이 경제활동인구조사와 차이를 보인다.

Table 2.1 Trend of the ratio of economically active population of women (2000-2016)

year	2000	2001	2002	2003	2004	2005	2006	2007	2008
ratio (%)	41.12	41.38	41.39	41.02	41.38	41.53	41.71	41.67	41.64
year	2009	2010	2011	2012	2013	2014	2015	2016	
ratio (%)	41.31	41.44	41.50	41.60	41.75	42.01	42.25	42.31	

한국노동연구원의「2015년 한국노동패널조사 (KLIPS)」데이터를 이용하여 국내 기혼여성의 임금에 미치는 요인을 파악하고자 한다. 한국노동패널조사는 1단계에서 조사구를 선정하고, 2단계에서 가구 선정하는 층화집락계통추출법을 사용하여 4,185가구를 추출하였다. 이 중 결측치가 있는 데이터를 제외하고, 개인 데이터에서 가구ID를 기준으로 가구주와 그 배우자를 매칭한 결과 총 3,357개 가구가 분석에 사용되었다. 기혼여성을 기준으로 분석되었으며, 결측치가 있는 데이터는 제외시켰다. 분석에 사용된 종속변수는 경제활동인구 여부와 작년 세전 근로소득이고, 독립변수는 나이, 사회보험 개수, 자격증 개수, 사회경제적 지위, 주당 평균 근로시간, 남편의 작년 세전 근로소득 등을 고려하였다. Table 2.2는 분석에 사용된 변수와 변수에 대한 설명이다.

Table 2.3은 기혼 여성의 기술통계 결과이다. 전체 3,357명 중에서 경제활동인구는 1,232명, 비경제활동인구는 2,125명이다. 비경제활동인구인 기혼여성의 경우는 평균 만5세 이하 자녀의 수가 많고, 부모님이나 자녀에게 금전적으로 도움을 받은 것으로 나타났다. 반면, 경제활동을 하고 있는 기혼여성의 경우는 평균 만 6세 이상 자녀의 수가 많고, 부모님이나 자녀에게 금전적으로 도움을 주는 것으로 나타났다. 또한, 평균 연령이 낮고, 여성이 교육을 받은 기간과 남편이 교육을 받은 기간이 높게 나타났으며, 최대 사회보험의 개수와 최대 자격증 개수가 많은 것으로 나타났다. 경제활동을 하고 있는 기혼여성의 가구는 생활비와 부채자산 총계의 평균이 높은 반면, 부채자산 총계의 최대 금액은 아내가 경제활동을 하지 않는 가구가 높은 것으로 나타났다. 이는 남편의 소득이 높은 가구인 것으로 확인되었다.

Table 2.2 Variables

	Variables	Description	Type	S*	O**
Independent variables	w_age	age	numeric	O	O
	w_edu	education period (year)	numeric	O	O
	w_ss	socioeconomic status : 1-lowest, 6-highest	ordinal		O
	w_worktime	average hours worked per week	numeric		O
	w_n_sinsu	the number of insurance	numeric		O
	w_n_certi	the number of certificates	numeric		O
	w_n_training	the number of training	numeric		O
	w_nchild1	number of children between 0 and 5 years old	numeric	O	
	w_nchild2	6 years - number of children under high school student	numeric	O	
	w_nchild3	number of university student (graduate student)	numeric	O	
	w_cost	last year's living expenses (10,000 won)	numeric	O	
	w_f_money1	the amount of money that receives from other people for economic help (10,000 won)	numeric	O	
	w_f_money2	the amount of money paid for other dependents (10,000 won)	numeric		O
	h_wage1	husband's last year's before tax earnings (10,000 won)	numeric	O	
	h_edu	husband's education period (year)	numeric	O	
Response variables	w_wage	Married women's income before tax (10,000 won)	numeric	O	
	w_ea	Economically active population : 0='nonactive', 1='active'	binary		O

S*: variables used in selection model, O**: variables used in outcome model

Table 2.3 Statistics for married women

Variables		Total			Economically active			Economically nonactive		
		Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
Independent variables	w_age	21	50.47	91	23	46.64	78	21	52.69	91
	w_edu	0	11.60	23	0	12.51	23	0	11.08	23
	w_ss	1	2.87	6	1	2.88	6	1	2.86	6
	w_worktime	0	15.06	100	1	41.04	100	0	0.00	0
	w_n_sinsu	0	0.05	2	0	0.04	2	0	0.06	1
	w_n_certi	0	0.01	4	0	0.02	4	0	0.01	2
	w_n_training	0	0.07	10	0	0.13	10	0	0.03	3
	w_nchild1	0	0.24	3	0	0.18	3	0	0.27	3
	w_nchild2	0	0.56	4	0	0.71	4	0	0.47	4
	w_nchild3	0	0.17	3	0	0.26	3	0	0.12	2
	w_cost	20	263.30	1,000	45	303.00	1,000	20	240.30	1,000
	w_f_money1	0	181.00	20,000	0	126.50	12,300	0	212.50	20,000
	w_f_money2	0	226.50	35,300	0	270.80	24,060	0	200.80	35,300
	h_wage1	0	3,278	36,000	0	3,370	19,200	0	3,225	36,000
	h_edu	0	12.63	23	0	13.15	23	0	12.32	23
Response variables	w_wage	0	827.2	10,000	33	2,033	10,000	0	128.3	6,000
	w_ea			3,357			1,232			2,125

3. 표본 선택 모형

3.1. Heckman 표본 선택 모형

Heckman 표본 선택 모형은 다음과 같이 두 개의 모형을 포함하고 있다.

$$y_i^* = x_i^T\beta + \epsilon_i, \quad i = 1, \ \cdots, \ N, \tag{3.1}$$

$$u_i^* = w_i^T\gamma + \eta_i, \quad i = 1, \ \cdots, \ N. \tag{3.2}$$

이 모형에서 주된 관심사는 여성의 임금에 영향을 주는 요인을 찾는 것이다. (3.1)에서 y_i^*은 여성의 임금, x_i는 임금 y_i^*에 영향을 주가지수는 요인이다. 여성은 임금은 육아, 경력 단절 등의 이유로 결측되는 경우가 종종 발생하므로, 표본 선택에 대한 모형을 고려할 필요가 있다. (3.2)는 이러한 표본 선택을 설명하는 모형이다.

$$u_i = I\ (u_i^* > 0),$$
$$y_i = y_i^* u_i.$$

위의 모형은 Tobit 모형 (Amemiya, 1984)의 특별한 경우이다. 모든 여성의 임금이 관측될 수 없으므로, 은 잠재 변수 (latent variable)로 볼 수 있다. 즉, 위의 모형에서 $u_i = 1$이면 y_i^*은 관측되고, $u_i = 0$이면 y_i^*은 관측되지 않는다. (3.1)로부터 다음과 같은 식을 얻을 수 있다.

$$E(y_i^*|x, u_i^* > 0) = x_i^T\beta + E(\epsilon_i|u_i^* > 0) = x_i^T\beta + E(\epsilon_i|\eta_i > -w_i^T\gamma).$$

따라서 $E(\epsilon_i|\eta_i > -w_i^T\gamma) = 0$이 아니면, 즉, ϵ과 η가 독립이 아닌 경우에는 (3.1)의 OLS 추정치 $\hat{\beta}$은 편의 (bias)를 갖게 된다. Heckman (1977)은 오차 ϵ_i과 η_i에 대해서 다음과 같이 이변량 정규분포 가정을 하였다.

$$\begin{pmatrix} \epsilon_i \\ \eta_i \end{pmatrix} \sim N_2(0, \ \Omega). \tag{3.3}$$

이 때, $\Omega = \begin{pmatrix} \sigma_1^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}$이다. 즉, ϵ과 η는 상관성을 가질 수 있는 구조를 고려하였다. 이러한 가정 하에서 x가 관측되었을 때 y의 평균은 다음과 같이 계산된다.

$$E\ (y \mid u = 1, \ x, \ w) = x^T\beta + \rho\sigma_1\lambda(w^T\gamma). \tag{3.4}$$

이 때, $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$은 inverse Mills ratio이다. Heckman (1976, 1977)은 (3.4)에 대해서 다음과 같이 2 단계 (two-step procedure)를 통해 모형을 추정하는 방법을 고안하였다.

1 단계: 프로빗 모형 (Probit model)을 이용하여 (3.2)에서 γ의 추정치 $\hat{\gamma}$을 구한다.
2 단계: (3.4)를 이용하여 주어진 $(x, \hat{\gamma})$에 대해서 $\hat{\beta}, \hat{\rho}, \hat{\sigma})$을 구한다.

이 방법은 Heckman (1974)에서 제시한 최대우도추정량을 이용한 방법보다 효율성이 떨어지나, 오차에 대한 정규분포 가정에 robust한 성질을 갖는다는 장점이 있다. 그러나 w가 (3.1)의 x의 일부를 포함하고 있을 때 다중공선성 문제가 발생할 수 있다. 따라서 오차에 대한 정규성 가정을 할 수 없고, x와 w간의 상관관계가 높을 때에는 Heckman (1974)이 제시한 최대우도추정 방법과 Heckman (1976,

1977)이 제시한 2 단계 모형 추정 방법은 효율성 및 일치성이 떨어질 수 있다. Heckman 2단계 모형은 최대 우도 추정 방법을 이용하지 않기 때문에 AIC와 BIC처럼 우도값에 근거한 모형 비교 및 모형 선택을 이용할 수 없다는 단점이 있다. Heckman 모형을 이용한 데이터 분석은 Stata와 R에서 sampleSelection 패키지를 이용할 수 있다.

3.2. Heckman 표본 선택 t 모형

Heckman (1977)은 (3.3)에서 이변량 오차에 대하여 정규분포 가정을 하였으나, 두꺼운 꼬리분포를 가정하는 것이 실제 데이터에 적합한 경우가 많다. Marchenko와 Genton (2012)은 이변량 오차에 대하여 다음과 같이 이변량 t 분포 가정을 하였다.

$$\begin{pmatrix}\epsilon_i\\ \eta_i\end{pmatrix} \sim t_2(0,\ \Omega,\ \nu). \tag{3.5}$$

이 때, $\Omega = \begin{pmatrix}\sigma_1^2 & \rho\sigma\\ \rho\sigma & 1\end{pmatrix}$이고, 이변량 t 분포 $t_2(0,\ \Omega,\ \nu)$는 다음과 같다.

$$f(t;\ \mu,\ \Omega,\ \nu) = (2\pi)^{-1}|\Omega\ |^{-1/2}\left\{1+\nu^{-1}(t-\mu)^T\Omega^{-1}(t-\mu)\right\}^{-(\nu+2)/2}. \tag{3.6}$$

$Y^*|U^* > 0$의 분포는 extended skew-t 분포 (Arellano-Valle *et al.*, 2010)로 다음과 같은 식으로 표현된다.

$$f_{EST}(y;\ \mu,\ \sigma^2,\ \alpha,\ \tau,\ \nu) = \frac{1}{\sigma}t(z;\nu)\frac{T[(\alpha z+\tau)((\nu+1)/(\nu+z^2)^{1/2});\nu+1]}{T(\tau/\sqrt{1+\alpha^2};\nu)}.$$

이 때, $z=(y-\mu)/\sigma$, $\alpha=\rho/\sqrt{1-\rho^2}$, $\tau = w^T\gamma/\sqrt{1-\rho^2}$, $T(\cdot;\nu)$는 $t(\cdot;\nu)$분포의 누적 확률 분포이다. 관측치의 조건부 기댓값은 다음과 같다.

$$E\ (y\mid u=1,\ x,\ w) = x^T\beta + \rho\sigma_1\lambda(w^T\gamma). \tag{3.7}$$

이 때, $\lambda_\nu(v) = \frac{\nu+v^2}{\nu-1}\frac{t(v;\nu)}{T(v;\nu)}$이다. (3.6)로부터 다음과 같은 로그 우도함수를 유도할 수 있다.

$$\begin{aligned} l(\beta,\ \gamma,\ \rho,\ \sigma,\ \nu\ ;\ y,\ u) =& u\ ln\ f(y\ |U=\ 1) + u\ ln\ T(w^T\gamma\ ;\ \nu) + (1-u)\ ln\ T(-w^T\gamma\ ;\ \nu)\\ =& u\ ln\ t(z\ ;\ \nu) - u\ ln\ \sigma + (1-u)\ ln\ T\ (-w^T\gamma;\nu)\\ &+ u\ ln\ T\ \left\{\ \left(\frac{\nu+1}{\nu+z^2}\right)^{1/2}\ \frac{\rho z + w^T\gamma}{\sqrt{1-\rho^2}}\ ;\ \nu+1\ \right\}. \end{aligned}$$

Marchenko와 Genton (2012)은 이러한 로그 우도함수를 최대로 하는 (β,γ)를 추정하는 방법을 제시하였다. 따라서 우도값에 근거한 AIC와 BIC를 계산하여 모형 선택에 이용할 수 있다. Heckman t 모형의 분석을 하기 위해서는 Marchenko와 Marc (2012)의 저자에게 요청하여 Stata에서 이용할 수 있는 코드를 제공받아서 이용할 수 있다. 다만, 이 코드에 AIC와 BIC를 계산하는 알고리즘이 아직 구현되어 있지 않다.

4. 데이터 분석

2절에서 소개한 데이터에 대해서 먼저 표본 선택 편의를 고려하지 않고, (3.1) 모형에 대해 OLS 모형으로 분석하고, 여성의 경제활동 여부에 대해서는 프로빗 모형을 이용해서 분석하도록 한다. OLS 회귀모형에서 경제 활동을 하지 않는 여성의 임금은 결측치로 간주하고, 모두 제외하고 분석하였다. 또한 3절에서 소개한 두 가지 표본 선택 모형, Heckman 2 단계 방법과 Heckman selection-t 방법을 적용하여 분석하였다. 식 (3.1)에 해당하는 주된 결과 모형 (outcome model)에서 종속변수는 여성의 임금으로 하고, 독립변수로는 나이 (w_age), 소유하고 있는 사회보험 수 (w_n_sinsu), 자격증 수 (w_n_certi), 사회경제적 지위 (w_ss), 주당 평균 근로시간 (w_worktime), 교육받은 기간 (w_edu), 교육 및 훈련 수 (w_n_training), 다른 사람으로부터 금전적으로 도움을 받은 금액 (w_f_money2)을 포함한 모형이다. 식 (3.2)에 해당하는 표본 선택 모형 (selection model)에서는 종속변수는 직업의 유무, 독립변수는 나이 (w_age), 교육받은 기간 (w_edu), 만 0세-만5세 자녀의 수 (w_nchild1), 만 6세-고등학생 (재수생) 자녀의 수 (w_nchild2), 대학(원)생 자녀의 수 (w_nchild3), 생활비 (w_cost), 다른 사람으로부터 금전적으로 도움을 받은 금액 (w_f_money1), 남편의 소득 (h_wage), 남편의 학력 (h_edu)을 고려하였다. 통계 분석을 위해서 Stata 14.0을 이용하였다. Table 4.1은 세 가지 방법에 해당하는 분석 결과이다.

각 변수에 대해 해석은 다음과 같이 할 수 있다. 결과 해석에 있어 계수의 값보다 부호에 의의를 두었다. 주된 결과 모형 (outcome model)에서의 공변량의 부호는 세 모형이 모두 일치함을 알 수 있다. 또한 계수 값도 큰 차이가 없어 보인다. 계수가 큰 차이가 없다는 것은, 식 (3.1)에서 ϵ과 η가 큰 상관관계가 갖고 있지 않다는 것을 의미한다. Table 4.1에서 Heckman 모형과 Heckman t 모형에서 $\hat{\rho}$은 각각 0.36, 0.24인 것을 고려하면, ϵ과 η의 상관성은 작다고 볼 수 있다. 하지만, 유의 확률은 조금 차이가 있다. Heckman t 모형에서는 부양가족에게 경제적 도움을 주는 금액(w_f_money2)이 유의 수준 10% 기준에서 유의하다고 볼 수 있으나, 다른 두 모형에서는 그렇지 않다. 분석 결과로부터, 여성의 나이가 적을수록, 사회보험의 수와 자격증의 수가 적을수록, 학력이 높을수록, 사회경제적 지위가 높을수록, 교육 및 훈련의 개수와 경제적으로 도움을 준 금액이 많을 때, 여성의 임금이 높은 것을 알 수 있다.

여성의 경제 활동 여부를 분석한 모형에서는 주된 모형 (outcome model)보다는 확실히 차이가 있었다. 프로빗 모형에서는 여성이 나이가 적을수록, 고등학생 이하의 자녀가 있을수록, 다른 사람으로부터 경제적 도움을 받는 금액이 적을수록, 남편의 임금이 적을수록, 남편의 학력이 낮을수록, 여성의 학력이 높을수록, 생활비가 많이 들수록 경제활동을 할 가능성이 높았다. Heckman 모형의 표본 선택 모형은, 여성의 나이가 적을수록, 학력이 낮을수록, 자녀의 연령대와 상관없이 자녀가 없는 경우, 남편의 임금이 낮을수록, 생활비가 많을수록, 경제적으로 도움을 받은 금액이 높을수록, 남편의 학력이 높을수록 여성이 경제활동을 가능성이 높게 나왔다. 그러나 Heckman t 모형에서는 대학(원)생 자녀의 유무의 계수가 다르게 나왔다. 즉, 대학(원)생 자녀가 있는 경우에 여성이 경제활동을 할 가능성이 높다고 해석할 수 있다. 각 변수의 유의확률은 OLS와 표본 선택 모형이 큰 차이가 남을 알 수 있다. 두 표본 선택 모형에서는 유의확률이 약간 차이가 있다. Heckman 모형에서는 남편의 임금이 유의수준 10%에서 유의한 변수이지만, Heckman t 모형에서는 그렇지 않다.

Table 4.1 Results of OLS regression with probit model and sample selection models

	OLS model			Heckman selection model			Heckman selection t-model		
Variable	Coef	p-value		Coef	p-value		Coef	p-value	
Outcome model									
w_age	-0.00821	0.00048	**	-0.00830	0.00035	**	-0.00711	0.00014	**
w_edu	0.06531	0.00000	**	0.06667	0.00000	**	0.06475	0.00000	**
w_ss	0.21934	0.00000	**	0.22500	0.00000	**	0.20691	0.00000	**
w_worktime	0.01844	0.00000	**	0.02527	0.00000	**	0.02097	0.00000	**
w_n_sinsu	-0.12441	0.17261		-0.14432	0.10766		-0.12943	0.11227	
w_n_certi	-0.35659	0.00058	**	-0.33814	0.00100	**	-0.24883	0.02672	**
w_n_training	0.11913	0.00232	**	0.12295	0.00153	**	0.12704	0.00021	**
w_f_money2	0.00002	0.23803		0.00002	0.29711		0.00004	0.05767	*
Probit model				Selection model					
w_age	-0.04323	0.00000	**	-0.03816	0.00000	**	0.05622	0.00000	**
w_edu	0.02950	0.01428	**	-0.02691	0.16075		-0.00979	0.77580	
w_nchild1	-0.58630	0.00000	**	-0.50017	0.00000	**	-0.69037	0.00001	**
w_nchild2	-0.06924	0.04387	*	-0.18547	0.00122	**	-0.29233	0.00292	**
w_nchild3	0.19475	0.00052	**	-0.04898	0.64538		0.00533	0.97500	
w_cost	0.00362	0.00000	**	0.00025	0.60272		0.00032	0.70093	
w_f_money1	-0.00006	0.11458		0.00003	0.53414		0.00008	0.10063	
h_wage1	-0.00019	0.00000	**	-0.00004	0.07505	*	-0.00005	0.14981	
h_edu	-0.03048	0.00491	**	0.00923	0.58860		0.00352	0.90751	
σ	-	-		0.71394	-		0.42045	-	
ρ	-	-		0.36369	-		0.24064	-	
ν	-	-		$+\infty$	-		2.84330	-	

Heckman 표본 선택 t 모형은 이변량 오차에 대하여 이변량 t 분포 가정을 한 것으로, 식 (3.5)에서 $\nu = \infty$인 경우 Heckmam 표본 선택 모형이 되므로, Heckman 표본 선택 t 모형이 더 일반화된 모형이라고 볼 수 있다. 표 4.1에서 $\hat{\nu} = 2.843$이므로, Heckman 표본 선택 모형보다는 Heckman 표본 선택 t 모형이 더 적합하다고 볼 수 있다. 따라서 Heckman 표본 선택 모형의 결과보다는 Heckman 표본 선택 t 모형의 결과를 받아들이는 것이 적합할 것으로 파악된다.

5. 결론

Marchenko와 Genton (2012)는 Heckman (1976, 1977)의 Heckman 표본 선택에서의 이변량 오차가 정규 분포를 따른다는 가정이 다소 강함을 지적하고 이변량 오차에 t 분포를 이용한 Heckman 표본 선택 t 모형을 제안하였다. 「2015년 한국노동패널조사 (KLIPS)」자료를 활용하여 국내 여성의 임금 결정요인을 분석한 결과, 주된 결과 모형에서는 추정치의 부호가 차이가 없었으나, 여성의 경제활동 여부에 영향을 주는 요인의 결과가 다소 다름을 알 수 있었다. Heckman 표본 선택 t 모형이 Heckman 표본 선택의 일반화된 모형임을 감안한다면, Heckman 표본 선택 t 모형에서 추정된 값이 크지 않을 경우, Heckman 표본 선택 t 모형의 결과를 해석하는 것이 바람직하다. 2015년 국내 여성의 경제 활동 여부에 영향을 주는 요인으로는 나이, 고등학생 이하 자녀의 수를 들 수 있고, 국내 여성의 임금에 영향을 주는 요인으로는 나이, 학력, 사회적 지위, 사회보험의 수, 자격증의 수, 교육 및 훈련 수, 다른 사람으로부터 경제적으로 도움 받는 금액을 들 수 있다. 이렇듯, 임금 결정 요인을 분석하는 모형으로 Heckman t 모형이 자주 이용될 수 있을 것으로 기대한다. 향후 연구로는, Ryu와 Cho (2016)이 분석한 특성화고 졸업생의 임금 결정요인에 대해서도 Heckman t 모형을 적용하고자 한다.

References

Amemiya, T. (1984). Tobit models: A survey. *Journal of Econometrics*, **24**, 3-61.

Arellano-Valle, R. B. and Genton, M. G. (2010). Multivariate extended skew-t distributions and related families. *Metron*, **68**, 201-234.

Marchenko, Y. V. and Marc, G. G. (2012). A Heckman selection-t model. *Journal of the American Statistical Association*, **107**, 304-317.

Heckman, J. J. (1974). Shadow prices, market wages, and labor supply. *Econometrica*, **42**, 679-694.

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. (1977). Sample selection bias as a specification error. *Econometrica*, **47**, 153-161.

Park, S. and Cho, J. (2016). The wage detrimentals applying sample selection bias. *Journal of the Korean Data & Information Science Society*, **27**, 1317-1325.

Ryu, J. and Cho, J. (2016). The wage determinants of the vocational high school graduates using mixed effects mode. *Journal of the Korean Data & Information Science Society*, **27**, 935-946.

Korea Labor Institute. `http://www.kli.re.kr`

Journal of the Korean Data & Information Science Society
2017, **28**(5), 1077–1085

http://dx.doi.org/10.7465/jkdi.2017.28.5.1077
한국데이터정보과학회지

Korean women wage analysis using selection models[†]

Mi Ryang Jeong[1] · Mijeong Kim[2]

[12]Department of Statistics, Ewha Womans University

Received 17 August 2017, revised 13 September 2017, accepted 14 September 2017

Abstract

In this study, we have found the major factors which affect Korean women's wage analysing the data provided by 2015 Korea Labor Panel Survey (KLIPS). In general, wage data is difficult to analyze because random sampling is infeasible. Heckman sample selection model is the most widely used method for analysing the data with sample selection. Heckman proposed two kinds of selection models: the one is the model with maximum likelihood method and the other is the Heckman two stage model. Heckman two stage model is known to be robust to the normal assumption of bivariate error terms. Recently, Marchenko and Genton (2012) proposed the Heckman selection-t model which generalizes the Heckman two stage model and concluded that Heckman selection-t model is more robust to the error assumptions. Employing the two models, we carried out the analysis of the data and we compared those results.

Keywords: Heckman selection model, Heckman selection-t model, missing not at random, sampling bias.

[†] Mijeong Kim was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean Government (NRF-2017R1C1B5015186).

[1] Graduate student, Department of Statistics, Ewha Womans University, 52, Ewhayeodae-gil, Seodaemun-gu, Seoul 03760 Korea.

[2] Corresponding author: Associate professor, Department of Statistics, Ewha Womans University, 52, Ewhayeodae-gil, Seodaemun-gu, Seoul 03760 Korea. E-mail: m.kim@ewha.ac.kr