The Korean Journal of Applied Statistics (2017) DOI: http://dx.doi.org/10.5351/KJAS.2017.30.6.917
30(6), 917–929

Bayesian analysis of Korean income data using zero-inflated Tobit model

Jisu Hwang[a] · Sei-Wan Kim[a] · Man-Suk Oh[b,1]

[a]Department of Economics, Ewha Womans University;
[b]Department of Statistics, Ewha Womans University

(Received September 5, 2017; Revised October 20, 2017; Accepted October 26, 2017)

Abstract

Korean income data obtained from Korea Labor Panel Survey shows excessive zeros, which may not be properly explained by the Tobit model. In this paper, we analyze the data using a zero-inflated Tobit model to incorporate excessive zeros. A zero-inflated Tobit model consists of two stages. In the first stage, individuals with 0 income are divided into two groups: genuine zero group and random zero group. Individuals in the genuine zero group did not participate labor market since they have no intention to do so. Individuals in the random zero group participated labor market but their incomes are very low and truncated at 0. In the second stage, the Tobit model is assumed to a subset of data combining random zeros and positive observations. Regression models are employed in both stages to obtain the effect of explanatory variables on the participation of labor market and the income amount. Markov chain Monte Carlo methods are applied for the Bayesian analysis of the data. The proposed zero-inflated Tobit model outperforms the Tobit model in model fit and prediction of zero frequency. The analysis results show strong evidence that the probability of participating in the labor market increases with age, decreases with education, and women tend to have stronger intentions on participating in the labor market than men. There also exists moderate evidence that the probability of participating in the labor market decreases with socio-economic status and reserved wage. However, the amount of monthly wage increases with age and education, and it is larger for married than unmarried and for men than women.

Keywords: Markov chain Monte Carlo, Tobit model, truncated data, zero-inflated data

1. 서론

우리나라 경제활동 참가율은 1980년대 60% 초반으로 OECD 평균보다 8% 가량 낮은 수치였으나 꾸준히 증가하여 2015년 기준 68.3%으로 OECD 평균과 차이가 약 2%대로 줄었다 (OECD labor force statistics, 2017). 생산가능인구인 만 15세부터 만 64세 인구 중 경제활동에 참여하는 인구의 비중인 경제활동참가율은 국가 생산성 성장의 중요한 요소로서 잠재경제성장률의 측도로 활용된다. 기존 문헌들

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the ministry of Education, Science and Technology (No. NRF-2016R1A2B4008914).

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

은 주로 경제활동참가율이 낮은 여성의 노동 공급 결정에 초점을 두었으나 본 논문에서는 남녀를 모두 포함한 데이터를 대상으로 분석하였다.

생산가능인구의 월평균 소득자료는 0보다 큰 범위에서는 연속적인 분포이나 0에서 절단된 형태를 보인다. 이처럼 종속변수가 한쪽에서 절단된(truncated) 경우 기존의 최소자승법으로 추정하는 것은 부적절하다. 경제활동에 참여하고 있는 개인들의 소득만 관측되고 그렇지 않은 경우 0으로 관측되어 표본 편의(sample bias)가 발생하기 때문이다. Tobin (1958)은 이를 해결하기 위해 전 구간에서 연속인 잠재변수를 도입한 토빗모형(Tobit model)을 설계하였다. Heckman (1979) 또한 취업 여부와 임금결정이 단계적으로 결정되는 2단계 모형(two-part model)을 이용하여 여성의 노동임금을 분석하였다. 토빗모형은 설명변수가 취업여부에 미치는 영향과 임금결정에 미치는 영향이 동일하다고 가정한 반면 Heckman의 2단계 모형은 본 논문과 유사하게 두 개의 다른 식을 적용하여 특정 설명변수가 노동시장 참여와 임금결정에 미치는 영향의 크기가 다를 수 있을 뿐만 아니라 영향을 미치는 요소 역시 다를 수 있다고 가정하였다.

절단된 자료에 대한 베이지안 연구도 꾸준히 진행되었다. Tanner와 Wong (1987)은 절단된 부분에 잠재변수를 도입하는 방법을 제안하였고 Gelfand와 Smith (1990)는 이를 바탕으로 모수의 완전 조건부 사후분포를 이용한 깁스 표본 기법(Gibbs sampling)을 소개하였다. 잠재변수를 활용한 깁스 표본 기법은 기존의 최우추정(maximum likelihood estimation; MLE), 라플라스 접근법과 비교하여 계산이 복잡하지 않으면서 모수를 잘 추정하였으며 0이 과도한 자료에서 다른 방법에 비해 0의 개수를 더 근접하게 추정하였다 (Chib, 1992; Ghosh 등, 2006).

본 논문에서 사용한 생산가능인구의 소득분포는 기존의 토빗모형에서 가정하는 0의 관측치보다 과도하게 많은 0이 포함된다 (Figure 4.1). 이러한 영과잉 자료에는 허들 모형(hurdle model)과 영과잉 모형(zero-inflated model)을 적용할 수 있다. 전자의 경우 영자료에는 이항분포를 가정하고 0보다 큰 자료에는 절단된 분포(truncated distribution)를 적용하여 두 분포에서 정의된 우도함수를 각각 최대화하는 방법으로 모수를 추정한다. (Cragg 등, 1971; Mullahy, 1986) 반면 후자는 영과잉 분포를 0에서 퇴화된 분포와 유한개의 0을 포함한 분포가 가중평균된 형태로 가정하여 각 우도함수의 가중평균을 최대화하는 모수를 사후추정치로 한다 (Lambert, 1992). 허들모형에서는 2단계에서 0자료가 발생하지 않으나 영과잉 모형의 경우 혼합된 각각의 분포에서 모두 0이 발생가능하다는 점에서 차이가 있다.

토빗모형을 이용하여 소득자료를 분석하면 0으로 나타난 부분에 소득분포와 동일한 분포를 따르는 잠재변수를 생성하므로 모든 개인이 시장에 참여할 의사가 있다고 가정한다. 또한 2단계 모형이나 허들모형의 경우 0자료를 노동시장 참여의사결정에서 나타난 하나의 이항분포에서 생성된 것이라 가정한다. 하지만 소득이 0으로 나타나는 노동시장 비참여는 원래부터 노동시장 참여의사가 없는 자발적 비참여와 노동시장에 참여하고자 했으나 낮은 시장임금으로 인해 참여를 포기한 비자발적 비참여가 섞여 있다. 이를 고려하여 본 논문은 영과잉 모형과 토빗모형의 아이디어를 기반으로 노동시장 참여의사여부와 0을 포함한 소득분포 결정으로 이루어진 영과잉 토빗모형을 사용하여 실제 소득분포를 분석하였다. Yang과 Simpson (2010)에 따르면 영과잉 토빗모형과 같은 계층적 모형은 1단계 의사 결정과 2단계 규모 결정을 구분된 단계로 추정하기 때문에 경계값의 확률과 나머지 분포에 대해 해석이 유연하다. 2단계에서 0 소득이 발생하지 않는 허들모형과 달리, 영과잉 토빗모형은 1단계와 2단계에서 0 소득이 발생하므로 노동시장 비참여를 자발성 여부에 따라 나눌 수 있을 뿐만 아니라 실제 소득은 0이지만 노동시장 참여의지가 있는 집합을 포함하여 임금결정모수를 추정할 수 있다.

노동시장 참여결정에는 성별, 교육 수준과 같은 개인적 특성 이외에도 사회적 요소가 영향을 미친다. 사회적 인적 네트워크를 통해 자신이 원하는 직장을 얻을 가능성이 높아 노동시장에 참여할 유인이 크기 때문이다 (Lin 등, 1981). 또한 토빗모형에 따르면 유보임금이 시장임금보다 낮은 경우 노동시장에

참여하므로 유보임금 수준이 노동시장 참여결정에 유의한 영향을 미친다고 가정할 수 있다. 본 논문에서는 내부적 요인인 1단계 노동시장 참여의사 여부에 사회경제적 지위와 자신이 생각하는 좋은 직장의 임금수준을 포함하여 두 변수의 영향을 검증하고자 하였다. 한국노동패널조사에서 제공하는 18차년도 Klips 개인용 설문자료를 바탕으로 만 15세부터 만 64세 표본의 월평균 소득분포를 이용하였다. 1단계 노동시장 참여의사여부에는 로짓분포를 적용하고 2단계 0이 포함된 소득분포에는 기존의 토빗모형을 이용하였다. 모수 추정은 메트로폴리스-헤스팅스(Metropolis-Hastings; M-H) 기법과 깁스(Gibbs) 표본 기법이 혼재된 M-H within Gibbs 기법을 이용하였다.

본 논문의 구성은 다음과 같다. 2절에서는 영과잉 토빗모형을 소개한다. 3장에서는 마코브 체인 몬테칼로(Markov chain Monte Carlo; MCMC)를 이용한 베이지안 분석방법을 소개한다. 4장에서는 실제 데이터를 활용하여 영과잉 토빗모형(zero-inflated Tobit model)으로 모수를 추정한 후 토빗모형과 비교하여 모형 적합성 결과를 제시하고 추정 결과를 해석하였다. 마지막 5장은 결론과 요약이다.

2. 모형

2.1. 토빗모형

반응변수 $Y_i (i = 1, 2, \ldots, n)$가 $[0, \infty)$에서 관측되며 0에서 점 확률을 갖고 $(0, \infty)$에서 연속분포를 지닌다고 가정하자. 종속변수가 0에서 절단된 경우 0보다 큰 Y_i만을 대상으로 모수를 추정하게 되면 편의가 발생한다. 이를 해결하기 위해 전 범위 $(-\infty, \infty)$에서 정의된 잠재변수 Y^*를 생성하여 Y^*가 0보다 큰 경우 Y^*값으로 관측되고 그렇지 않은 경우 0으로 관측되는 모형을 생각할 수 있다 (Tobit, 1958).

$$y_i^* = X_i'\beta + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right), \tag{2.1}$$

$$f\left(y_i|\beta, \sigma^2\right) = \begin{cases} P(y_i^* < 0), & \text{if } \ y_i = 0, \\ f\left(y_i^*|\beta, \sigma^2\right), & \text{if } \ y_i > 0. \end{cases} \tag{2.2}$$

식 (2.1)에서 X_i를 $k \times 1$ 설명변수 벡터라고 할 때, β는 $k \times 1$ 모수 벡터이다. ϵ/σ는 대칭인 표준정규분포를 따르므로 $P(y_i^* < 0) = \Phi\left(-X_i'\beta/\sigma\right)$이다. 식 (2.2)로부터 $y = (y_1, \ldots, y_n)$의 밀도함수는

$$f\left(y|\beta, \sigma^2\right) = \prod_{i=1}^{n} \Phi\left(-\frac{X_i'\beta}{\sigma}\right)^{I(y_i=0)} \times \phi\left(\frac{X_i'\beta}{\sigma}\right)^{I(y_i\neq 0)} \tag{2.3}$$

이다. 또는 잠재변수 $y^* = (y_1^*, \ldots, y_n^*)$를 고려하면

$$f\left(y, y^*|\beta, \sigma^2\right) = \prod_{i=1}^{n} \phi\left(\frac{y_i^* - X_i'\beta}{\sigma}\right) \left[I(y_i^* < 0, y_i = 0) + I(y_i^* > 0, y_i = y_i^*)\right] \tag{2.4}$$

이다.

2.2. 영과잉 토빗모형

0이 매우 과다하게 관측되는 자료의 경우 토빗모형을 이용하여 자료를 적합시키는 것은 한계가 있을 수 있다. 영과잉 토빗모형은 0 관측값이 2가지 요인으로부터 유래되었다고 가정한다. 첫 번째 요인은 0에서 점확률을 갖는 원조제로(genuine zero)이고 두 번째 요인은 토빗모형에서 절단된 형태의 0 관측값이다. 다시 말하면 원조제로에서 기인한 0 관측값에 토빗모형에서 기인한 0이 추가되어 과도한 0 관측값이 발생한다는 것이다. 영과잉 토빗모형의 분포를 표현하면

$$f_z\left(y_i|\beta, \sigma^2\right) = \pi_i \delta_0(y_i) + (1 - \pi_i) f\left(y_i|\beta, \sigma^2\right) \tag{2.5}$$

이다. π_i는 원조제로의 확률이며 $\delta_0(y_i)$는 $y_i = 0$일 때 1을 갖는 점확률, $f(y_i|\beta, \sigma^2)$는 식 (2.2)에 주어진 토빗모형의 밀도함수이다.

원조제로확률 π_i에 공변량을 도입하면

$$\text{logit}(\pi_i(\gamma)) = \log\left(\frac{\pi_i(\gamma)}{1-\pi_i(\gamma)}\right) = Z_i'\gamma \tag{2.6}$$

의 연결함수를 고려할 수 있다. 이 때 공변량 Z_i는 X_i와 같을 수도 있고 다를 수도 있다.

3. 베이지안 분석

MCMC를 이용한 베이지안 분석을 용이하게 하기 위하여 잠재변수 $S_i \sim \text{Ber}(\pi_i)$를 모형에 도입하면 식 (2.5)는

$$\begin{aligned} f_z\left(y_i, S_i|\beta, \gamma, \sigma^2\right) &= [\delta_0(y_i)]^{S_i}\left[f(y_i|\beta,\sigma^2)\right]^{1-S_i}\pi_i(\gamma)^{S_i}(1-\pi_i(\gamma))^{1-S_i} \\ &= [\pi_i(\gamma)\delta_0(y_i)]^{S_i}\left[(1-\pi_i(\gamma))f(y_i|\beta,\sigma^2)\right]^{1-S_i} \end{aligned} \tag{3.1}$$

로 나타낼 수 있다. 또한 식 (2.4)를 이용하면

$$\begin{aligned} &f\left(y_i, y_i^*, S_i|\beta,\gamma,\sigma^2\right) \\ &= [\pi_i(\gamma)\delta_0(y_i)]^{S_i}\left[(1-\pi_i(\gamma))\phi\left(\frac{y_i^* - X_i'\beta}{\sigma}\right)\left\{I(y_i^* \le 0, y_i = 0) + I(y_i^* = y_i, y_i > 0)\right\}\right]^{1-S_i} \end{aligned} \tag{3.2}$$

이다.

모수 β, γ, σ^2의 사전분포로

$$\begin{aligned} \beta &\sim \mathcal{N}(\beta_0, \Sigma_0), \\ \gamma &\sim \mathcal{N}(\gamma_0, R_0), \\ \sigma^2 &\sim \mathcal{IG}(a, b) \end{aligned}$$

을 가정한다. $\mathcal{IG}(a,b)$는 평균이 $b/(a-1)$인 역감마분포를 나타낸다. 식 (3.2)의 우도함수와 위의 사전분포로부터 잠재변수와 모수의 조건부 사후분포를 유도하면 다음과 같다. 단, S_i의 사후분포 유도에는 식 (3.1)의 우도함수를 사용하였다.

$$\begin{aligned} \beta|\text{else} &\propto \prod_{i:S_i=0}\phi\left(\frac{y_i^* - X_i'\beta}{\sigma}\right)\pi(\beta) \\ &\sim \mathcal{N}\left(\left(\frac{1}{\sigma^2}X_0'X_0 + \Sigma_0^{-1}\right)^{-1}\left(\frac{1}{\sigma^2}X_0'y_0^* + \Sigma_0^{-1}\beta_0\right), \left(\frac{1}{\sigma^2}X_0'X_0 + \Sigma_0^{-1}\right)^{-1}\right), \end{aligned}$$

여기에서 X_0는 $S_i = 0$에 해당하는 X_i들만 모은 $n_0 \times k$ 행렬이고, y_0^*는 $S_i = 0$에 해당하는 y_i^*들만 모은 n_0차원 벡터이며, n_0는 $S_i = 0$인 i들의 개수다.

$$y_i^*|\text{else}, S_i = 0 \sim \begin{cases} \mathcal{N}\left(X_i'\beta, \sigma^2\right)I(y_i^* < 0), & \text{if } \ y_i = 0, \\ \mathcal{N}\left(X_i'\beta, \sigma^2\right)I(y_i^* > 0), & \text{if } \ y_i > 0, \end{cases}$$

$$\gamma|\text{else} \propto \prod_{i=1}^{n}\pi_i(\gamma)^{S_i}\left((1-\pi_i(\gamma))^{1-S_i} \times \pi(\gamma),\right.$$

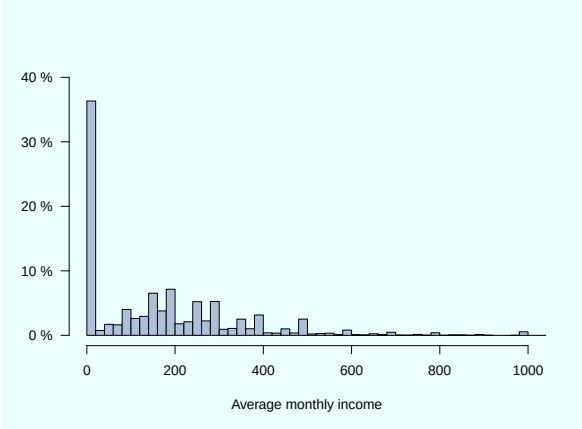


Figure 4.1. Average monthly income distribution.

$$S_i|\text{else} \sim \begin{cases} \text{Ber}\left(\dfrac{\pi_i(\gamma)}{\pi_i(\gamma)+(1-\pi_i(\gamma))\Phi\left(-\frac{X_i'\beta}{\sigma}\right)}\right), & \text{if } \ y_i = 0, \\ \delta_0(S_i), & \text{if } \ y_i > 0, \end{cases}$$

$$\sigma^2|\text{else} \sim \mathcal{IG}\left(\frac{n_0}{2}+a, \frac{1}{2}(y_0^*-X_0\beta)'(y_0^*-X_0\beta)+b\right).$$

γ를 제외한 나머지 변수들의 조건부 사후분포가 편리한 형태로 주어지므로 이들의 조건부 사후분포로부터 사후표본을 생성하고 γ의 경우 랜덤워크 메트로폴리스 기법을 수행하여 사후표본을 생성하는 M-H within Gibbs 기법을 적용한다.

4. 한국인의 소득자료 분석

4.1. 데이터 설명

본 논문은 한국노동패널조사에서 제공하는 18차년도 Klips 개인용 설문자료에 영과잉 토빗모형을 적용하여 분석하였다. 설문자료는 2015년에 총 14,012명을 대상으로 취직 여부 및 임금 분포 데이터를 제공하고 있다. 우리나라는 의무 교육 제도와 높은 대학진학률로 인하여 사회진출이 상대적으로 늦다. 만 15세부터 만 18세는 생산가능인구에 속하지만 소득이 거의 0으로 나타나기 때문에 원조제로확률 추정 시, 나이효과가 지배적이 될 가능성이 높다. 이러한 문제를 방지하기 위해 본 논문에서는 사회진출이 활발한 만 19세부터 정년인 만 64세의 표본 10,197명을 대상으로 월평균 소득자료를 분석하였다.

원래부터 노동시장 참여의지가 없는 확률, 즉 원조제로확률을 설명하는 변수로는 성별(gen), 결혼 여부(mar), 학력 수준(edu), 나이(age), 사회경제적 지위(prop), 자신이 생각하는 좋은 직장의 월평균 최소 임금 수준(rwage)을 사용하였다. 사회경제적 지위는 자신의 소득, 직업, 교육, 재산을 모두 고려하여 주관적으로 평가한 지표로서 하하(1), 하상(2), . . . , 상하(5), 상상(6) 총 6단계로 구분된다. 자신이 생각하는 좋은 직장의 임금 수준은 자신이 원하는 일자리를 얻기 위한 주관적인 척도로 유보임금의 대리변수로 사용하였다. 시장에서 정해지는 임금보다 유보임금이 상당히 높은 경우 노동시장 참여의지에 영향을 미칠 수 있으므로 1단계 설명변수로 포함하였다. 반면 2단계 노동시장임금 분석에서는 참여의지가 있는 사람만을 대상으로 하므로 사회경제적 지위변수와 유보임금 대리변수를 제외한 성별, 결혼 여부, 학력 수준, 나이를 설명변수로 하여 모형을 구성하였다.

Table 4.1. The estimates for average monthly income data by two models

Model	Parameter	Variable	Estimate (SE)
Zero-inflated Tobit model (M-H within Gibbs)	participation(π)	gen(γ_1)	-2.3977 (0.2420)
		mar(γ_2)	-2.5690 (1.3373)
		edu(γ_3)	1.9079 (0.1943)
		age(γ_4)	-7.7253 (0.5277)
		prop(γ_5)	0.2338 (0.1077)
		rwage(γ_6)	-0.4247 (0.2117)
	wage(μ)	gen(β_1)	-1.2881 (0.0269)
		mar(β_2)	0.2004 (0.0326)
		edu(β_3)	0.3078 (0.0151)
		age(β_4)	0.0403 (0.0180)
	variance	σ^2	1.4185 (0.0265)
Tobit model	wage(μ)	gen(β_1)	-1.1981 (0.0271)
		mar(β_2)	0.4395 (0.0331)
		edu(β_3)	0.3185 (0.0156)
		age(β_4)	0.2091 (0.0177)
	variance	σ^2	1.5868 (0.0296)

SE = standard error; M-H = Metropolis-Hastings.

Table 4.2. Comparison two models for the data

	The number of zeros	Sums of squared error
True data	3605	-
Zero-inflated Tobit model	3337	8739.249
Tobit model	2748	8919.500

4.2. 추정 방법 및 결과

γ와 β의 사전분포는 사전정보의 영향을 배제하기 위하여 분산이 큰 다변량 정규분포 $\mathrm{MVN}(0, 10^2 I)$로 선택하였다. γ의 초기치는 양수인 y를 1로 변환하여 로짓분포로 일반화 선형 회귀모형에서 얻은 추정치로 설정하고 β는 단순 선형 회귀의 추정치로 설정하였다. σ^2의 사전분포는 모수가 $(a, b) = (0.5, 1)$인 역감마분포로 설정하고 초기치는 주어진 데이터의 표본분산으로 가정하였다. 앞 장의 M-H within Gibbs 알고리즘을 이용하여 각 모수의 사후표본을 3개의 Chain으로 15만 개 생성한 후에 수렴된 표본만을 사용하기 위해 5만 번은 제거(burn in)하였다. 각 모수의 Gelman 통계량이 모두 1.1을 넘지 않고 경로 그림을 확인한 결과 사후표본이 잘 수렴한다고 보인다. Table 4.1은 추정결과이며 Figure 4.2에서 각 모수의 사후 밀도 함수와 95% 최대사후구간(highest posterior density interval; HPD)을 확인할 수 있다. 모수 γ_2의 사후밀도함수는 왼쪽으로 긴 꼬리를 가진(skewed) 분포를 보여 최고 사후 최빈값(mode)을 사후추정치로 하였다. 나머지 모수들의 사후밀도함수는 좌우대칭형태로 최고 사후 최빈값과 사후표본의 평균이 유사하였다.

4.3. 모형의 적합성

Table 4.2에 따르면 영과잉 토빗모형이 토빗모형보다 실제 0개수에 근접하게 추정하였고 오차제곱합(sums of squared error; SSE)도 토빗모형에 비해 작은 것을 확인할 수 있다. 영과잉 토빗모형이 기존 토빗모형에 비해 주어진 데이터에 대해 얼마나 더 적합한지 알아보기 위하여 Bayesian information criteria (BIC), deviance information criteria (DIC), mean logarithm conditional predictive ordinate

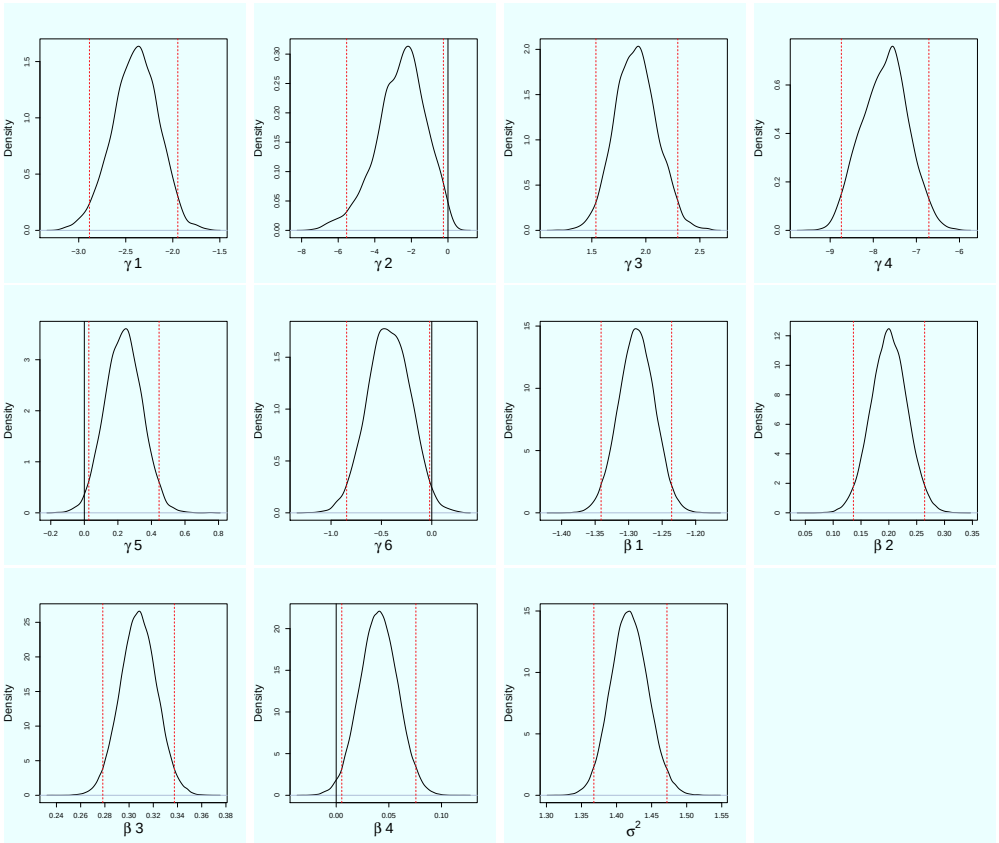


Figure 4.2. Posterior density plots for parameters. 95% HPD intervals are given with dotted lines and the solid line is at 0.

($\overline{\text{LCPO}}$) (Meng, 1994), 그리고 posterior predictive p-value (PPP) (Carlin과 Louis, 2009)를 계산하였다 (Table 4.3). 토빗모형에 비하여 영과잉 토빗모형의 BIC, DIC, $\overline{\text{LCPO}}$이 작고 또한 PPP가 0.5에 더 가까우므로 상대적으로 영과잉 토빗모형이 데이터에 더 적합하다고 판단할 수 있다.

4.4. 추정 결과 해석

원조제로확률인 π_i는 경제활동에 참가하지 않는 비율(economically inactive rate)을 의미한다. 본 논문에서는 사회진출이 활발한 만 20세부터 만 64세만을 대상으로 추정했기 때문에 경제활동 참가율인 $(1-\hat{\pi}_i)$의 평균값은 생산가능인구(만 15세–만 64세)를 대상으로 한 경우(약 68%)보다 훨씬 높은 값(약 93%)으로 나타났다. 영과잉 토빗모형에 따르면 성별, 결혼 여부, 교육수준, 나이, 사회경제적 지위, 유보임금이 노동시장 참여의사에 유의한 영향을 미치며 자신이 생각하는 좋은 직장의 최소임금은 95%에서 유의한 것으로 나타났다. 각 추정치의 부호에 따르면 남성일수록, 미혼일수록, 교육수준이 높을수록, 나이가 적을수록, 주관적 사회경제적 지위가 높을수록, 유보임금이 낮을수록 원래부터 노동시장에 참여할 의지가 없을 확률이 높은 것으로 나타났다. 특히 나이의 영향이 가장 높고 결혼 여부와 성별이 그 뒤를 이었다. 사회경제적 지위와 유보임금의 유의성은 상대적으로 강하지 않다. 구체적으로 한계효과와 각 변수별 영향을 분석하면 다음과 같다.

Table 4.3. Fitness of two models for the data

	BIC	DIC	$\overline{\text{LCPO}}$	PPP
Zero-inflated Tobit model	26404.41	26308.86	2.5125	0.4436
Tobit model	27293.07	27249.66	2.6271	0.7127

BIC = Bayesian information criteria; DIC = deviance information criteria; $\overline{\text{LCPO}}$ = mean logarithm conditional predictive ordinate; PPP = posterior predictive p-value.

Table 4.4. Average marginal effect

Variable	Average marginal effect	Variable	Average marginal effect
gen	−0.0477	age	−0.1536
mar	−0.0517	prop	−0.0046
edu	0.0377	rwage	−0.0090

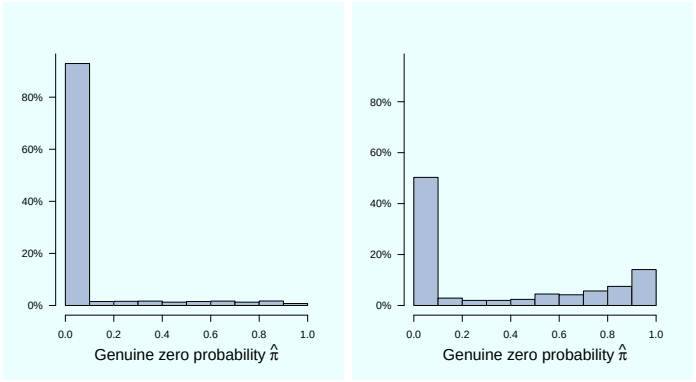
4.4.1. 한계효과 Table 4.4는 식 (4.1)로 정의된 평균 한계 효과(average marginal effect)를 이용한 것으로 각 변수별 한계효과를 나타낸 것이다. 변수가 한 단위 늘어날 경우 원조제로확률에 미치는 영향을 평균한 값이다. 나이의 평균 한계 효과는 −0.1546으로 나이가 한 단위 늘어날수록 원조제로확률이 15.5% 낮아진다. 한편 결혼여부의 한계효과는 −0.0514로 기혼일 때 미혼일 때보다 원조제로확률이 평균적으로 5.1% 낮아지며 남성의 경우 여성에 비해 원조제로 확률이 평균적으로 4.7% 낮다.

- 평균 한계 효과

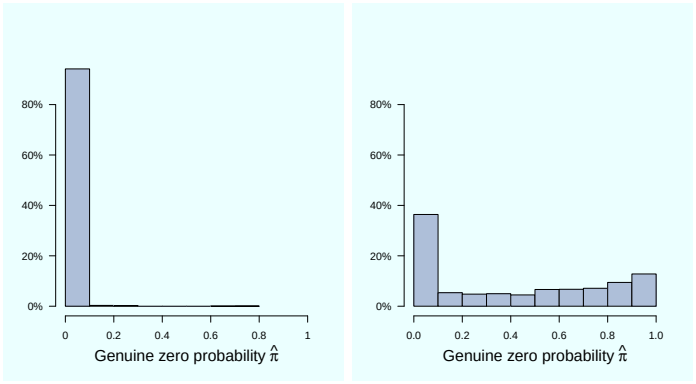
$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\frac{\partial \pi_i}{\partial z_{ij}} &= \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial z_{ij}}\left(\frac{e^{\gamma_0+\gamma_1 z_{i1}+\cdots+\gamma_q z_{iq}}}{1+e^{\gamma_0+\gamma_1 z_{i1}+\cdots+\gamma_q z_{iq}}}\right)\\ &= \frac{1}{N}\sum_{i=1}^{N}\gamma_j\frac{e^{Z'\gamma}}{(1+e^{Z'\gamma})^2} = \frac{1}{N}\sum_{i=1}^{N}\gamma_j\pi_i(1-\pi_i), \quad j\in\{1,2,\ldots,q\}. \end{aligned} \tag{4.1}$$

4.4.2. 사회경제적 지위가 노동시장 참여의지에 미치는 영향 사회경제적 지위는 개인의 물질적 풍요 상태와 주관적 만족을 종합적으로 나타낸 지표로서 건강상태와 연계되어 노동시장 참여와 퇴출에 유의한 영향을 미칠 수 있다 (Schuring 등, 2013). 사회경제적 지위가 낮을수록 육체적, 정신적으로 미약하여 노동시장에서 저소득을 받거나 극단적으로 시장에서 퇴출될 가능성이 높기 때문이다. 또한 사회경제적 지위는 개인의 물질적 조건 이외에 사회적 자원인 인적 네트워크를 포함하므로 사회경제적 지위가 높을수록 노동시장에서 자신이 원하는 직업을 얻을 가능성이 높다 (Lin 등, 1981). 하지만 이는 표본 모두가 노동시장에 참여할 의사가 있다고 가정하고 사회경제적 지위의 영향을 살펴본 것으로 본 논문은 사회경제적 지위가 노동시장 참여의사여부 자체에 미치는 영향을 살펴보았다. 사후추정치는 0.2338으로 사회경제적 지위가 높아질수록 원조제로확률에 양(+)의 영향을 미치는 것으로 나타났다. 즉, 자신이 생각하는 사회경제적 지위가 높을수록 애초에 노동시장에 참여할 의지가 낮을 수 있다는 것이다. 이러한 추정결과의 의미는 사회경제적 지위가 노동시장에 참여할 유인을 제공하지 못함을 의미한다.

4.4.3. 유보임금이 노동시장 참여의지에 미치는 영향 유보임금(reservation wage)이란 노동경제학에서 노동자가 시장에 참여할 최소한의 임금 수준을 의미한다. Jones (1988)은 실제자료를 이용하여 유보임금이 실업 기간에 양의 영향을 미친다는 것을 보였다. 하지만 이는 노동시장 참여의지가 있는 개인들만을 대상으로 분석한 것이다. 이와 달리 본 논문은 유보임금이 노동시장 참여의지 여부에 미치는 영향을 추정하였다. 유보임금은 주관적 수치로서 관측이 불가능하므로 본 논문에서는 유보임금의 대리변



(a) Female with zero income (b) Male with zero income



(c) Married with zero income (d) Not married with zero income

Figure 4.3. Distribution of genuine zero probability.

수로 '자신이 생각하는 좋은 직장의 최소 임금' rwage을 사용하였다. 'rwage'는 엄밀히 말하면 고용되기 위한 최소임금은 아니다. 하지만 타인과 비교 시, 상대적인 위치는 유지하므로(ordinal) 유보임금과 평행적인 차이(level effect)만 있다고 가정할 수 있다. 사후추정치에 따르면 'rwage'가 높을수록 노동시장에 참여할 의지가 있는 것으로 나타났다. 95% 사후 최고 밀도 구간의 경계값이 0에서 멀지는 않지만 'rwage'는 원조제로확률을 낮추는 것으로 추정되었다. 이는 유보임금이 높을수록 노동시장에 참여할 의지가 있을 가능성이 높다는 것을 의미한다.

4.4.4. 성별이 노동시장 참여 및 소득분포에 미치는 영향 성별 변수는 여성인 경우 1이고 남성인 경우 0으로 처리하였다. 총 표본 중 소득이 0인 3,605명만을 고려한 결과 여성은 대부분 노동시장 참여의지가 높은 반면 남성은 노동시장 참여의지가 낮은 비중이 여성에 비해 큰 것으로 나타났다. Figure 4.3(a)를 보면, 소득이 0인 여성은 총 2,649명으로 원조제로확률이 거의 0에 가깝게 추정되고 약 14%인 370명만이 원조제로확률이 0.1에서 1사이에 고르게 분포하였다. 즉, 소득이 0인 여성의 경우 대다수가 노동시장에 참여할 의지가 매우 높으나 개인의 특성인 성별, 결혼 여부, 교육, 나이에 따라 결정된 시장임금(market wage)이 자신의 유보임금보다 높지 않아 노동시장에 참여하지 못한 것이다. 반면 소득이 0인 남성은 총 956명으로 약 50%만 노동시장에 참여할 의사가 확실하였고 나머지는 노동시장에 참여

할 의지가 불분명하였다. Figure 4.3(b)에 따르면 남성의 원조제로확률은 0과 1에서 봉우리를 갖는 양봉(bimodal) 분포를 보인다. 즉, 여성은 주로 외부적 요인에 의해 노동시장 비참여가 발생하는 반면 남성은 외부적 요인과 내부적 요인이 모두 작용하여 노동시장 비참여가 발생한다고 볼 수 있다. 이는 노동시장에서 나타나는 성별 직업 분리와 성별 간 임금 격차를 반영하는 결과라고 볼 수 있다. 여성의 사회참여율은 꾸준히 증가하여 2000년대부터는 50%대로 남성의 참여율과 격차가 줄어들고 있으나 아직도 사회통념상 성별에 따라 여성직, 남성직으로 직업군을 분리하는 경우가 많다. 여성이 접근 가능한 직장은 고용이 불안정하고 저소득인 경우가 많으며 직업 선택의 폭 역시 남성에 비해 좁은 실정이다 (Keum, 2011). 나아가 성별 직업 분리는 임금 분포에서 성별이 미치는 음의 효과를 부분적으로 설명한다. 고소득 직업에 남성이 포진해 있는 반면 여성은 주로 저소득 직업에 분포한 것을 반영한 것이라 판단된다. 성별 간 임금 격차 또한 여성의 낮은 임금 형성에 일조한다는 점에서 노동시장 참여에 미치는 영향을 설명할 수 있다.

4.4.5. 결혼 여부가 노동시장 참여 및 소득에 미치는 영향 결혼 여부는 미혼과 사별인 경우 0으로 하고 이혼, 기혼, 별거의 경우 기혼으로 1로 처리하였다. 결혼을 하게 되면 가족의 생계유지라는 경제활동에 대한 유인이 생기므로 노동시장 참여의사에 유의한 영향을 미치는 요소라고 예측할 수 있다. 추정결과, 결혼여부는 노동시장 참여의사에는 음(−)의 영향을 주고 소득에는 양의 영향을 주는 것으로 나타났다. Figure 4.3의 (c)–(d)를 보면, 소득이 0인 기혼자의 경우 거의 대부분 노동시장 참여의사가 있으나 노동시장조건에 의해 시장에 참여하지 못한 것으로 추정되었다. 한편 소득이 0인 미혼자는 상대적으로 0과 1사이에 고르게 분포하였다. 즉 과다하게 추정된 0자료는 자발적으로 노동시장에 참여하지 않는 미혼자 그룹에서 기인한 것을 알 수 있다. 미혼인 경우 부모와 함께 살고 있거나 떨어져 살아도 기혼자에 비해 외부의 재정적 지원이 가능하므로 노동시장에 참여할 의사가 상대적으로 낮다고 해석할 수 있다. 한편 노동시장 참여의사가 있는 개인들만을 분석한 결과 계수 추정치가 0.2004로 결혼 여부가 소득에 양(+)의 영향을 미치는 것으로 나타났다. 이는 결혼 프리미엄(marital premium)을 반영한 결과로 여성의 경우 결혼의 효과가 불분명하지만 남성의 경우 기혼남성의 소득이 미혼남성보다 유의하게 높다. 기혼자의 경우 결혼을 통한 심리적 안정감과 가정에 대한 책임으로 인한 생산성 향상이 존재하기 때문이다 (Korenman와 Neumark, 1991).

4.4.6. 나이가 노동시장 참여 및 소득에 미치는 영향 나이 변수는 19세 부터 64세까지를 표준정규화하여 데이터에 포함하였다. 기존 토빗모형에 따르면 나이는 소득분포에 유의한 양(+)의 효과를 미치는 것으로 나타났으나 영과잉 토빗모형에 의하면 나이는 1단계 노동시장 참여의지와 2단계 소득분포 결정에 상반된 영향을 갖는 것으로 나타났다. 먼저 나이의 사후추정치는 −7.7253로 노동시장 참여의지에 가장 유의한 음(−)의 영향을 미쳤다. 이는 나이가 늘어날수록 노동시장에 참여할 의지가 높다는 것을 의미하며 그 영향의 크기가 지배적이다. 한편 원조제로가 아닌 경우 즉, 노동시장에 참여의사가 있는 경우에는 나이가 다른 변수들에 비해 소득분포에 유의한 영향을 미치지 않는 것으로 나타났다. 일단 노동시장에 참여하고자 하는 의지가 있으면 더이상 나이의 영향이 크지 않다는 것으로 노동시장 참여결정을 단계별로 나누지 않으면 알 수 없는 부분이다. 또한 이는 나이에 따라 소득이 꾸준히 증가하다가 40대 이후 점차 감소하여 나이 효과가 상쇄되는 현실 상황과 일치하는 결과이다 (Saint-Pierre, 1996).

5. 결론

흔히 0에서 절단된 소득자료는 절단 회귀 모형인 토빗모형이나 Heckman의 2단계 모형으로 분석한다. 하지만 실제 소득자료에서 기존의 토빗모형이 가정하는 0자료 수보다 과도한 0이 관측된다. 이를 고

려하여 본 논문은 0에서 퇴화한 분포와 토빗분포의 혼합분포인 영과잉 토빗모형을 이용하여 실제 소득분포를 분석하였다. 모수 추정은 메트로폴리스-헤스팅스 샘플링과 깁스샘플링을 혼합한 M-H within Gibbs 샘플링을 이용하였으며 모형 비교를 위해 동일한 자료에 대해 기존의 토빗모형으로도 모수를 추정하였다. 영과잉 토빗모형은 기존 토빗모형에 비해 0의 개수를 잘 예측하였으며 오차도 작았다. 또한 계층적 모형의 적합성 측도인 DIC를 비교한 결과 영과잉 토빗모형이 해당 데이터에 더 적합하다는 것을 확인할 수 있다.

본 논문은 소득자료에 영과잉 토빗모형을 적용하여 소득이 0인 대상을 내부적 요인에 의한 것인지 외부적 요인에 의한 것인지 구분하여 해석하였다. 노동시장 비참여를 원래부터 노동시장에 참여할 의사가 없는 자발적 비참여와 노동시장에서 정해진 낮은 임금으로 인한 비자발적 비참여로 나누었다. 추정 결과 사회경제적 지위가 높을수록, 유보임금이 낮을수록 노동시장 참여의사가 없을 가능성이 높았다. 소득이 0인 개인을 대상으로 본 결과 여성이거나 기혼인 경우 노동시장 참여의지는 높으나 시장의 외부적 요인에 의한 비자발적 노동시장 비참여가 높았다. 반면 미혼이거나 남성인 경우 원래부터 노동시장에 참여할 의지가 없어 소득이 0인 경우가 상대적으로 많았다. 나이 변수의 영향을 보면, 나이가 노동시장 참여의사에 매우 강한 양의 효과를 보이지만 노동 시장 참여시 소득에 대한 양의 효과는 상대적으로 크지 않았다.

References

Carlin, B. P. and Louis, T. A. (2009). *Bayesian Methods for Data Analysis* (3rd ed), Chapman & Hall/CRC, Boca Raton.

Chib, S. (1992). Bayes inference in the Tobit censored regression model, *Journal of Econometrics*, **51**, 79–99.

Cragg, J. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods, *Econometrica*, **39**, 829–844.

Gelfand, A. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association*, **85**, 398–409.

Ghosh, S. K., Mukhopadhyay, P., and Lu, J. C. (2006). Bayesian analysis of zero-inflated regression models, *Journal of Statistical Planning and Inference*, **136**, 1360–1375.

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

Jones, S. (1988). The relationship between unemployment spells and reservation wages as a test of search theory, *Quarterly Journal of Economics*, **103**, 741–765.

Keum, J. H. (2011). A study on the stagnation of the gender wage differences in Korea, *Kukje Kyungje Yongu*, **17**, 161–184.

Korenman, S. and Neumark, D. (1991). Does marriage really make men more productive?, *The Journal of Human Resources*, **26**, 282–307.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing, *Technometrics*, **34**, 1–14.

Lin, N., Vaughn, J., and Ensel, W. (1981). Social resources and occupational status attainment, *Social Forces*, **59**, 1163–1181.

Meng, X. L. (1994). Posterior predictive p-values, *Annals of Statistics*, **22**, 1142–1160.

Mullahy, J. (1986). Specification and testing of some modified count data models, *Journal of Econometrics*, **33**, 341–365.

Saint-Pierre, Y. (1996). Do earnings rise until retirement?, *Perspectives on Labour and Income*, **8**, 32–36.

Schuring, M., Robroek S. J., Otten F. W., Arts C. H., and Burdorf A. (2013). The effect of ill health and Socio economic status on labor force exit and re-employment: a prospective study with ten years follow-up in the Netherlands, *Scandinavian Journal of Work, Environment and Health*, **39**, 134–143.

Tanner, M. and Wong, W. (1987). The calculation of posterior distributions by data augmentation, *Journal of the American Statistical Association*, **82**, 528–540.

Tobin, J. (1958). Estimation of relationships for limited dependent variables, *Econometrica*, **26**, 24–36.

Yang, Y. and Simpson, D. G. (2010). Conditional decomposition diagnostics for regression analysis of zero-inflated and left-censored data, *Statistical Methods in Medical Research*, **21**, 393–408.

영과잉 토빗모형을 이용한 한국 소득분포 자료의 베이지안 분석

황지수[a] · 김세완[a] · 오만숙[b,1]

[a]이화여자대학교 경제학과, [b]이화여자대학교 통계학과

(2017년 9월 5일 접수, 2017년 10월 20일 수정, 2017년 10월 26일 채택)

요 약

한국노동패널조사에서 제공하는 2015년 한국 생산가능인구의 월평균 소득분포를 보면 0 관측치의 비율이 과도하게 높은 형태를 보여 기존의 소득분포에 주로 사용되는 토빗모형으로는 설명에 한계가 있다. 본 연구에서는 영과잉 특성을 반영하여 영과잉 토빗모형을 사용하여 한국인의 소득 자료를 분석한다. 영과잉 토빗모형은 2단계 모형으로 1단계에서는 소득이 0인 그룹을 두 그룹으로 나누는데, 첫 번째 그룹은 노동시장 참여의지가 없어 시장에 참여하지 않으므로 0이 관측되는 그룹(genuine zero)이고 두 번째 그룹은 노동시장 참여의지는 있으나 낮은 임금으로 인하여 절단되어 0이 관측되는 그룹(random zero)으로 가정하였다. 두 번째 random zero 그룹은 0 이상의 연속 자료와 결합하여 토빗모형을 적용한다. 1단계와 2단계 모형에 관심 있는 설명변수를 가진 회귀모형을 적용하여 노동시장 참여여부와 임금 수준에 영향을 미치는 요인을 알아본다. 마코브 체인 몬테칼로 기법을 사용하여 모수를 추정하고 기존의 토빗모형과 비교한 결과 영과잉 토빗모형이 0의 빈도추정과 모형 적합도 면에서 우수한 결과를 보였다. 분석 결과 나이가 많을수록, 남자가 여자보다, 학력이 낮을수록, 노동시장에 참여할 가능성이 매우 유의하게 높으며, 사회경제적 지위가 높을수록 그리고 유보임금이 낮을수록 노동시장에 참여하지 않을 확률이 높은 것으로 나타났다. 임금 수준을 보면, 남자가 여자보다, 학력이 높을수록, 기혼이 미혼 보다 매우 유의하게 더 높은 임금을 받는 것으로 나타났다.

주요용어: 마코브 체인 몬테칼로 기법, 영과잉 자료, 절단된 자료, 토빗모형

이 논문은 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 중견연구사업임 (No. NRF-2016R1A2B4008914).

[1]교신저자: (03760) 서울특별시 서대문구 이화여대길 52, 이화여자대학교 통계학과. E-mail: msoh@ewha.ac.kr