The Korean Journal of Applied Statistics (2013) DOI: http://dx.doi.org/10.5351/KJAS.2013.26.5.737
26(5), 737–746

Bayesian Interval Estimation of Tobit Regression Model

Seung-Chun Lee[a,1] · Byung Su Choi[b]

[a]Department of Applied Statistics, Hanshin University
[b]Department of Multimedia Engineering, Hansung University

(Received July 22, 2013; Revised October 21, 2013; Accepted October 21, 2013)

Abstract

The Bayesian method can be applied successfully to the estimation of the censored regression model introduced by Tobin (1958). The Bayes estimates show improvements over the maximum likelihood estimate; however, the performance of the Bayesian interval estimation is questionable. In Bayesian paradigm, the prior distribution usually reflects personal beliefs about the parameters. Such subjective priors will typically yield interval estimators with poor frequentist properties; however, an objective noninformative often yields a Bayesian procedure with good frequentist properties. We examine the performance of frequentist properties of noninformative priors for the Tobit regression model.

Keywords: Gibbs sampling, noninformative prior, censored regression model, coverage probability.

1. 서론

임상실험의 생존분석 데이터에는 중도절단된 관측값이 포함되어 있는 경우가 흔히 있다. 일반적으로 선형모형의 모수는 최소제곱법에 의해 추정되고 있으나 이와 같이 중도절단된 데이터에 의한 최소제곱추정량은 편의추정량(biased estimator)이라는 것이 잘 알려져 있으며 특히, 중도절단된 데이터의 비중이 클 경우, 편의의 정도가 커지기 때문에 설명변수의 영향을 과소, 또는 과대 추정하게 된다. 그러므로 최소제곱법의 의한 분석은 오류를 피할 수 없다. 이러한 이유로 Amemiya (1984), Green (1990) 등 많은 학자들이 반응변수가 중도절단된 토빗모형(Tobit model)에서 모수 추정방법을 연구하였다.

토빗회귀모형에 대한 빈도학파의 추정방법은 최대가능도추정법(maximum likelihood estimation)으로 귀결되며, `censReg` 패캐지의 `censReg` 함수, `VGAM` 패키지의 `Tobit` 함수, `survival` 패키지의 `survreg` 등 다양한 패키지에 구현되어 있다. 한편, Chib (1992)은 베이지안 추정에서 모수의 사후분포 기대값을 몬테칼로 적분, 라플라스 근사법 (Tierney와 Kadane, 1986) 및 깁스샘플링(Gibbs sampling)으로 구하는 방법에 대해 모의실험을 하였다. 이 모의실험에서 그는 표본크기가 큰 경우는 베이지안 추정과 최대가능도추정은 큰 차이를 보이지 않지만, 표본크기가 작은 경우에는 베이지안 추정값이 최대가능도 추정값보다 실제값에 가까운 것으로 결론지었다. 그러나 베이지안 추정값의 표준오차가 최대가능도 추정의 표준오차보다 모든 표본크기에서 대체로 큰 값을 갖는다고 하였다. 이는 두 추정방법 중 어느 하

This work was supported by Hanshin University research grant.
This research was financially supported by Hansung University.
[1]Corresponding author: Professor, Department of Applied Statistics, Hanshin University, 441 Yangsan-Dong, Osan, Kyunggi-Do, 447-791, Korea. E-mail: seung@hs.ac.kr

나의 표준오차는 실제값보다 과대 또는 과소 추정이 되었음을 의미한다. 본 연구의 동기는 최대가능도 추정의 표준오차가 과소추정이 되었을 가능성에 대해 알아 보려는 것이다. 즉, 최대가능도 추정량의 표준오차는 때때로 과소 또는 과대 추정이 되어 정규근사에 의한 Wald 신뢰구간의 포함확률(coverage probability)이 명목 신뢰수준과 유의한 차이를 보이는 경우 (Lee, 2006)가 있기 때문에 최대가능도 추정에 의한 구간추정의 유효성을 베이지안 구간추정과 비교하여 살펴보려고 한다.

Chib (1992)은 토빗모형의 베이지안 추정에서 깁스샘플링에 의한 추정이 매우 효율적이었다고 기술하고 있다. 그러나 베이지안 추정의 효율은 일반적으로 사후 기대값을 구하는 방법보다는 사전분포(prior distribution)에 의존하는 바가 크다. 특히, 구간추정에 있어서 사전분포는 추정의 효율에 크게 영향을 미친다. 즉, 베이지안 방법론에서 사전분포는 분석자의 사전정보를 반영하게 되는데 이와 같이 주관적인 분석은 빈도학파적 입장에서의 효율성을 담보하기 어렵다. 무정보사전분포(noninformative prior)에 의한 베이지안 추론에서는 분석결과가 사전정보 보다는 데이터 자체에 의존하는 바가 크므로 빈도학파적 효율성을 갖는 경우가 많다. 본 연구에서는 토빗모형에서 무정보사전분포에 의한 베이지안 구간추정의 효율성을 최대가능도추정방법과 비교하여 살펴보기로 한다.

2. 토빗모형

2.1. 최대가능도추정법

반응변수 $y_i^*, i = 1, \ldots, n$에 대한 회귀모형

$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N\left(0, \sigma^2\right) \tag{2.1}$$

에서 y_i^*가 구조적으로 중도절단되어 다음과 같이 y_i으로 관찰되었다고 하자.

$$y_i = \begin{cases} u, & i \in c_u, \\ y_i^*, & i \in c_m, \\ v, & i \in c_l, \end{cases}$$

여기서 $c_u = \{i : y_i^* \geq u\}$, $c_m = \{i : l \leq y_i^* \leq u\}$, $c_l = \{i : y_i^* \leq l\}$이다. 이와 같은 회귀모형에서 로그우도함수는

$$\ell\left(\boldsymbol{\beta}, \sigma^2\right) = \ell(\boldsymbol{\theta}) = \sum_{i \in c_m} \ln\left[\frac{1}{\sigma}\phi\left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right] + \sum_{i \in c_u} \ln\left[\Phi\left(\frac{\mathbf{x}_i'\boldsymbol{\beta} - u}{\sigma}\right)\right] + \sum_{i \in c_l} \ln\left[\Phi\left(\frac{l - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right]$$

와 같이 나타낼 수 있고, 최대가능도추정값과 추정 분산은 로그가능도방정식(log-likelihood equations)과 정보행렬(information matrix)

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}, \quad -\mathrm{E}\left(\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\right)$$

을 이용하여 구할 수 있다. 로그가능도방정식(log-likelihood equations)과 정보행렬(information matrix)의 자세한 식은 Henningsen (2012)을 참조하기 바란다.

앞에서 언급한 바와 같이 토빗모형의 최대가능도추정값은 `R` 패캐지인 `censReg`을 이용하여 구할 수 있으며, 또한 상용 통계 패키지인 `Stata`도 이용가능하다. 특히 `Stata`에서는 회귀계수에 대한 신뢰구간을 출력하고 있어 구간추정이 필요한 경우 편리하게 이용할 수 있다. 이때 신뢰구간은 회귀계수의 추정값과 표준오차를 이용하여 Wald 신뢰구간을 구하여 준다. 물론 이와같은 신뢰구간은 `censReg`의 출력물을 이용하면 손쉽게 구할 수 있다.

2.2. 베이지안추정

$\pi(\boldsymbol{\theta})$를 모수 $\boldsymbol{\theta}$에 대한 사전분포라고 하면, 베이지안 방법론에서 $\boldsymbol{\theta}$는 사후기대값

$$\mathrm{E}(\boldsymbol{\theta}|\mathbf{y}) = \frac{\int \boldsymbol{\theta} L(\boldsymbol{\theta};\mathbf{y})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int L(\boldsymbol{\theta};\mathbf{y})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{2.2}$$

으로 추정된다. 여기서 $L(\boldsymbol{\theta};\mathbf{y})$은 가능도함수(likelihood function)를 나타내는데, 사전분포가 가분포(improper distribution)인 경우 식 (2.2)의 분모항이 유한이면 사후기대값이 존재하는 것으로 알려져 있다.

일반적으로 베이지안 추론에서는 식 (2.2)의 값을 해석적(analytic) 방법으로 구하지 못하는 경우가 많기 때문에 이를 해결하기 위한 여러 가지 수치적 방법론이 연구되어 왔다. Chib (1992)은 왼쪽에서 절단된(left-censored) 토빗모형에서 사후기대값을 구하기 위해 몬테칼로 적분, 라플라스 근사 및 깁스샘플링 방법을 적용해 보고 모의실험을 통해 깁스샘플링 방법에 의한 추정이 상대적으로 효율성이 높다고 하였다. 깁스샘플링 방법은 사후 기대값을 구하는 과정에서 사후분포에 대한 추론이 가능하므로 구간추정에 사용하기 용이하다는 장점도 있다. 이러한 이유로 토빗모형의 베이지안 추정을 위한 대부분의 `R` 패캐지들은 깁스샘플링 방법을 채택하고 있고, 본 연구에서도 모수들의 구간추정을 위해 깁스샘플링 방법을 적용하기로 한다.

깁스샘플링을 위한 완전조건부 분포(full conditional distribution)들을 구하기 위해 $v_i, i \in c_l$과 $u_i, i \in c_u$를 각각 왼쪽과 오른쪽에서 절단된 잠재변수(latent variable)라고 하자. 이때 $v_i|\boldsymbol{\beta},\sigma^2$는 $(-\infty, l\,]$에서 독립적으로 절단정규분포를 따른다. 마찬가지로 $u_i|\boldsymbol{\beta},\sigma^2$도 $[\,u,\infty)$에서 절단정규분포를 따른다. 즉,

$$f\left(v_i|\boldsymbol{\beta},\sigma^2,\mathbf{y}\right) = \frac{\frac{1}{\sigma}\phi\left(\frac{v_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\frac{l - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)}, \quad v_i \in (-\infty, l\,],$$

$$f\left(u_i|\boldsymbol{\beta},\sigma^2,\mathbf{y}\right) = \frac{\frac{1}{\sigma}\phi\left(\frac{u_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\frac{\mathbf{x}_i'\boldsymbol{\beta} - u}{\sigma}\right)}, \quad u_i \in [\,u,\infty)$$

가 성립되며, 깁스샘플링의 i-번째 표본추출 과정은 다음과 같이 진행된다.

단계 1. $f(v_i|\boldsymbol{\beta}^{(i-1)},\sigma^{2(i-1)},\mathbf{y})$와 $f(u_i|\boldsymbol{\beta}^{(i-1)},\sigma^{2(i-1)},\mathbf{y})$에서 각각 $\mathbf{v}^{(i)} = (v_i^{(i)}, i \in c_l)$와 $\mathbf{u}^{(i)} = (u_i^{(i)}, i \in c_u)$을 추출.

단계 2. $\pi(\boldsymbol{\beta}|\,\mathbf{v}^{(i)},\mathbf{u}^{(i)},\sigma^{2(i-1)},\mathbf{y})$에서 $\boldsymbol{\beta}^{(i)}$ 추출.

단계 3. $\pi(\sigma^2|\,\mathbf{v}^{(i)},\mathbf{u}^{(i)},\boldsymbol{\beta}^{(i)},\mathbf{y})$에서 $\sigma^{2(i)}$ 추출.

깁스샘플링 과정에서 $\boldsymbol{\beta}$와 σ^2의 조건부 분포는 사전분포에 따라 다르다. `R`에서 토빗모형의 베이지안 추정을 위한 대표적인 함수로는 `MCMCpack` 패키지의 `MCMCtobit` 함수가 있고 이밖에 `Zelig` 패키지의 함수 `zelig`에서도 옵션을 통해 깁스샘플링을 통한 베이지안 추정을 할 수 있다. 이러한 함수들은 사전분포로써 $\boldsymbol{\beta}$와 σ^2의 준 켤레분포(semi-conjugate prior)으로 알려진 정규분포와 감마분포를 이용하고 있다. 즉,

$$\boldsymbol{\beta} \sim N\left(\mathbf{b}_0, \mathbf{B}_0^{-1}\right), \qquad \sigma^{-2} \sim \mathrm{Gamma}\left(\frac{c_0}{2}, \frac{d_0}{2}\right)$$

Table 3.1. Estimates and confidence intervals of Tobit model for Tobin data.

	Maximum likelihood			Semi-conjugate prior			Noninformative prior		
	Estimate(se)	95% CI		Estimate(se)	95% CI		Estimate(se)	95% CI	
intercept	15.145(16.08)	−16.37	46.66	18.249(41.22)	−59.15	100.2	17.207(37.37)	−53.34	92.87
age	−0.129(0.219)	−0.557	0.299	−0.281(0.607)	−1.603	0.660	−0.271(0.553)	−1.506	0.621
quant	−0.046(0.058)	−0.160	0.069	−0.048(0.149)	−0.336	0.229	−0.044(0.136)	−0.312	0.223
logSigma	1.718(0.310)	1.109	2.326	2.298(0.493)	1.488	3.411	2.256(0.471)	1.482	3.347

가 이용된다. $\mathbf{b}_0, \mathbf{B}_0^{-1}, c_0, d_0$는 사전정보에 의해 분석자가 정할 수 있는 값이지만 특정한 정보가 없는 경우 디폴트값인 $\mathbf{b}_0 = 0, \mathbf{B}_0^{-1} = 0, c_0 = 0.001, d_0 = 0.001$으로 무정보에 가까운 사전분포를 사용할 수 있도록 되어 있다. 여기서 $\mathbf{B}_0^{-1} = 0$의 설정은 $\boldsymbol{\beta}$에 대한 사전분포가 균등분포(uniform distribution)임을 의미한다. 한편, 회귀모형에서 일반적인 무정보사전분포는

$$\pi\left(\boldsymbol{\beta}, \sigma^2\right) \propto \frac{1}{\sigma^2} \tag{2.3}$$

이며 Chib (1992)도 깁스샘플링을 위해 식 (2.3)을 이용하고 있다.

식 (2.1)의 모형에서 관찰값 $\mathbf{y}$는 절단된 관찰값과 절삭없이 관찰된 $\mathbf{y}_1$으로 구성되어 있다. 절단된 관찰값에 대응하는 잠재변수를 $\mathbf{z} = (\mathbf{u}, \mathbf{v})$라고 하면 $\pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}) = \pi(\boldsymbol{\theta}|\mathbf{y}_1, \mathbf{z})$임을 쉽게할 수 있다. 이와 같은 사실을 이용하면 식 (2.3)의 사전분포에 대한 $\boldsymbol{\beta}$와 σ^2의 완전조건부 분포는

$$\begin{aligned}
&\boldsymbol{\beta}|\ \mathbf{v}, \mathbf{u}, \sigma^2, \mathbf{y} \sim N\left(\hat{\boldsymbol{\beta}}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right), \\
&\sigma^2|\ \mathbf{v}, \mathbf{u}, \boldsymbol{\beta}, \mathbf{y} \sim \text{InvGamma}\left(\frac{n}{2}, \frac{1}{2}(\mathbf{y}_z - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y}_z - \mathbf{X}\boldsymbol{\beta})\right)
\end{aligned} \tag{2.4}$$

과 같이 유도할 수 있다. 단 $\mathbf{y}_z = (\mathbf{y}_1, \mathbf{z})$이고, $\hat{\boldsymbol{\beta}}$는 $\boldsymbol{\beta}$의 최소제곱추정량 $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_z$이다.

Chib (1992)은 비록 한쪽으로 절단된 토빗모형에서 깁스샘플링을 위한 조건부 분포를 제시하였으나 이 결과는 식 (2.4)와 동일하다. 식 (2.4)의 자세한 유도과정은 Lee와 Lee (2002)를 참조할 수 있다.

3. 실증자료 분석

3.1. Tobin 데이터

Tobin 데이터는 2개의 변수 age, quant(liquidity ratio), durable(durable goods purchase)에 대한 20개의 관찰값으로 구성된 간단한 데이터로서 Tobin (1958)이 토빗모형을 처음 소개하면서 사용된 데이터이다. 반응변수 durable은 13개의 관찰값이 0에서 왼쪽으로 절단이 되어 있기 때문에 최소제곱법에 의한 추정은 편의가 심하게 나타날 뿐 아니라 표본크기가 작기 때문에 추정의 효율성이 매우 낮을 수 밖에 없다. 그러나 Chib (1992)은 표본크기가 작을 경우 최대가능도 추정과 베이지안 추정의 차이가 크다고 하였으므로 두 추정방법을 비교하기에 적합한 데이터로 판단된다.

Tobin 데이터의 추정 결과는 Table 3.1과 같다. 최대가능도 추정값과 준결레사전분포에 의한 베이지안 추정은 각각 `R`의 `censReg`와 `MCMCtobit` 함수로부터 구하였고, 무정보사전분포에 의한 추정은 `R` 함수를 작성하여 계산하였다. 표에서 최대가능도 추정의 신뢰구간은 정규근사에 의한 Wald 신뢰구간이고 베이지안 신뢰구간은 사후분포의 2.5 백분위점과 97.5 백분위점에서 구하였다.

Figure 3.1은 사후분포의 커널밀도함수추정(kernel density estimate)을 모수의 추정값과 같이 나타낸 것인데, 이 그림에서 추정값의 차이를 잘 보여 주고 있다. 즉, Tobin 데이터의 경우 표본크기가 작음

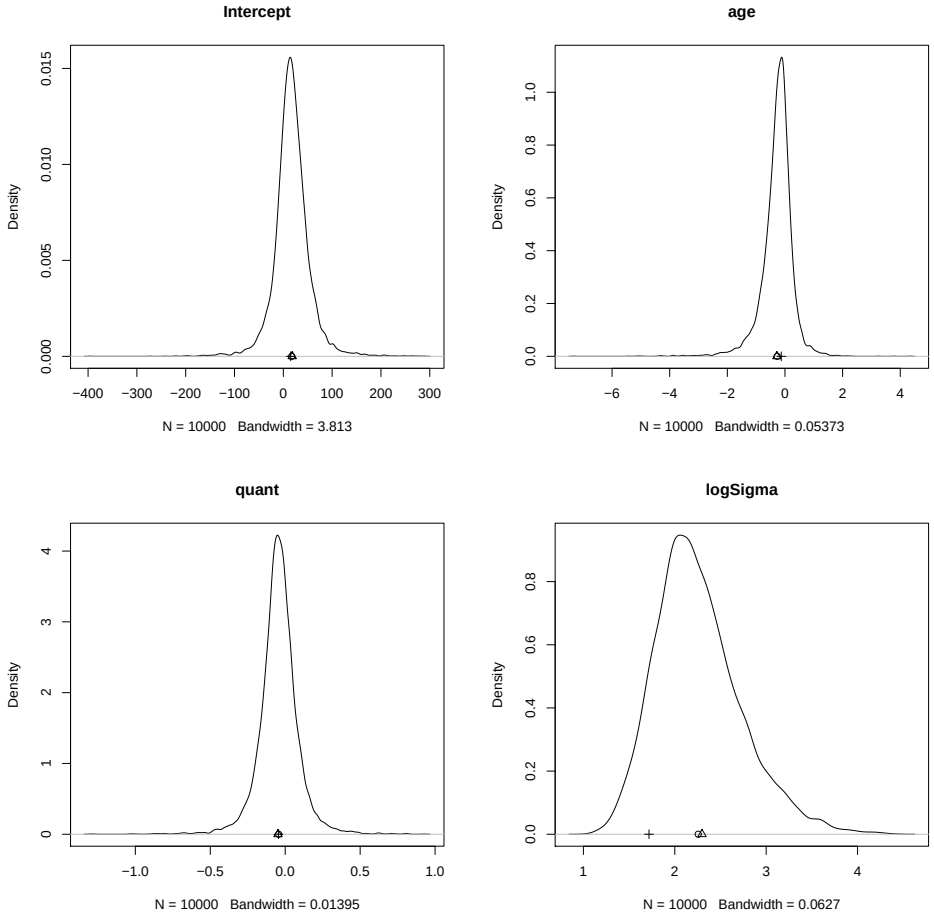

Figure 3.1. Estimates(MLE("+"), semi-conjugate("△"), noninformative("◯")) and Posterior distribution of parameters

에도 불구하고 logSigma를 제외한 회귀계수들의 추정값은 추정방법별로 큰 차이를 보이지 않는다. 그러나 표에서 최대가능도 추정의 추정오차는 베이지안 추정의 추정오차보다 약 2∼3배의 크기이므로 차이가 크다고 할 수 있고, 특히 절편의 추정오차 차이는 심각하다. 이와 같은 추정오차의 차이는 신뢰구간의 넓이에 영향을 미치게 되어 회귀계수에 대한 최대가능도 신뢰구간은 베이지안 신뢰구간보다 매우 짧은 것을 볼 수 있다. 그러므로 최대가능도 신뢰구간의 포함확률(coverage probability)이 명목신뢰수준(nominal confidence level)에 못미치거나, 또는 베이지안 신뢰구간이 매우 보수적인 신뢰구간일 수 밖에 없다. 두 추정방법에서 표준오차의 차이는 오차항의 분산 σ^2의 추정값 차이에서 나타나는 것으로 보여진다. 즉, σ^2의 최대가능도 추정값은 31.062인 반면, 베이지안 추정에서는 각각 186.846과 160.731로 추정된 바, 이 차이가 회귀계수들의 표준오차 계산에 영향을 미친 것으로 판단된다.

한편, 회귀계수들의 사후분포는 좌우대칭에 가까우므로 베이지안 추정량의 정규근사를 이용한 Wald 유형의 신뢰구간도 유용할 것으로 생각된다. 이와 같은 신뢰구간은 백분위점에 의한 신뢰구간보다 계산의 간편성에서 뛰어 나며 전통적인 베이지안 구간추정 방법인 최고사후밀도(Highest Posterior Density)

Table 3.2. Estimates and length of confidence interval of Tobit model for Morz data.

	Maximum likelihood estimate			Semi-conjugate prior estimate			Noninformative prior estimate		
	Estimate(se)	95% CI		Estimate(se)	95% CI		Estimate(se)	95% CI	
Intercept	965.31(446.4)	90.307	1840.3	953.97(450.8)	73.057	1834.6	958.50(448.4)	86.602	1826.2
nwifeinc	−8.814(4.459)	−17.55	−0.075	−8.780(4.521)	−17.82	−0.074	−8.878(4.508)	−17.73	−0.093
educ	80.646(21.58)	38.343	122.95	81.478(22.01)	39.014	125.00	81.428(21.83)	39.080	124.09
exper	131.56(17.28)	97.697	165.43	133.13(17.41)	99.717	168.07	132.76(17.51)	97.977	167.41
exper2	−1.864(0.538)	−2.918	−0.810	−1.895(0.546)	−2.963	−0.821	−1.879(0.543)	−2.954	−0.820
age	−54.41(7.419)	−68.95	−39.87	−54.84(7.514)	−69.84	−40.41	−54.88(7.482)	−69.49	−40.45
kidslt6	−894.0(111.9)	−1113.	−674.7	−901.3(115.1)	−1131.	−676.6	−901.5(113.0)	−1122.	−684.3
kidsge6	−16.22(38.64)	−91.95	59.518	−15.53(39.47)	−92.04	62.705	−15.38(38.85)	−91.67	60.780
logSigma	7.0229(0.037)	6.9503	7.0955	7.0361(0.038)	6.9638	7.1108	7.0361(0.038)	6.9660	7.1139

구간을 사후분포의 정규근사를 통해 구한 것으로 볼 수 있기 때문에 베이지안의 입장에서도 선호된다고 하겠다. 그러나 오차항의 분산 σ^2의 사후분포는 오른쪽으로 치우친 분포(right skewed distribution)이므로 정규근사에 의한 HPD 구간이 효용성을 갖기 어렵다. 오차항 분산의 경우 정규성 근사를 위해 σ에 대한 로그변환을 시도하였으나 표본크기가 작은 경우 Figure 3.1에서와 같이 비대칭적인 사후분포를 따르고 있었다. 따라서 σ^2 또는 $\log\sigma$의 구간추정에서는 HPD 보다 백분위점에 의한 신뢰구간이 적절한 것으로 보인다.

3.2. Morz 데이터

Morz 데이터는 결혼한 백인 여성 753명의 노동시간(working hour, hour), 여성의 나이(age), 교육년수(educ), 임금(wage), 6세 이하의 자녀 수(kidslt6), 6세 이상 19세 이하의 자녀 수(kidge6), 부인 임금 이외의 가구 소득(nwifeinc) 등이 수록된 자료로서 결혼한 여성의 근로참여 여부에 대한 로지스틱 또는 프로빗 모형의 예제로서 흔히 사용되고 있는 잘 알려진 데이터이다. 그러나 데이터의 출처인 Morz (1987)에서는 노동시간에 영향을 미치는 요인을 분석하기 위해 반응변수 hour에 대한 회귀모형에 사용되었다. 즉, Morz의 분석에서는 근로 시간이 0보다 큰 428명 자료로써 반응변수에 영향을 미치는 요인을 분석하였다. 따라서 Morz의 분석에서는 근로를 하지 않고 있는 여성 325명에 대한 분석이 결여 되어 있다고 볼 수 있다.

비근로여성에 대한 자료를 회귀모형에 포함할 경우 비 근로여성의 노동시간이 0에서 절단된 토빗모형이 적절하며 Wooldridge (2009)에 토빗모형에 대한 최대가능도 추정값이 수록되어 있다. 여기에서는 Wooldridge (2009)에 수록된 토빗모형에 대해 최대가능도추정과 준결레분포 및 무정보사전분포에 의한 베이지안 추정을 비교하기로 한다. 추정된 회귀계수와 추정오차 및 회귀계수들의 95% 신뢰구간은 Table 3.2과 같다.

Table 3.2에서 세 추정방법의 추정값은 큰 차이를 보이지 않고 있다. 추정값의 표준오차도 베이지안 추정방법이 최대가능도추정법과 비교하여 약간씩 크지만 값의 차이는 크지 않았다. 이러한 결과는 Chib (1992)이 모의실험 결과와 다르지 않다. 즉, Chib은 베이지안 추정에 의한 표준오차가 대체로 큰 경향이 있다고 하였고, 표본크기가 큰 경우 최대가능도추정값과 베이지안 추정값이 거의 같다고 언급하였다. 최대가능도 추정의 표준오차가 베이지안 추정의 표준오차보다 약간씩 큰 이유는 Tobin 데이터의 분석에서와 마찬가지로 오차항 분산의 추정과 관련이 있다. 즉, Figure 3.2에서와 같이 $\log\sigma$의 최대가능도 추정값이 큰 차이는 아니지만 베이지안 추정값보다 작았으며 이 결과가 회귀계수들의 표준오차 계산에 영향이 미친 것으로 생각된다.

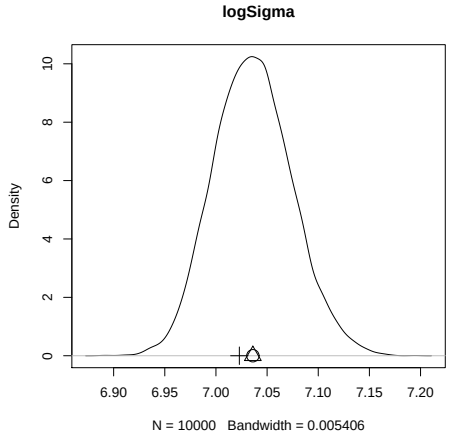

Figure 3.2. Estimates(MLE("+"), semi-conjugate("△"), noninformative("◯")) and Posterior distribution of $\log\sigma$

결론적으로 표본크기가 큰 경우 최대가능도 추정과 베이지안 추정은 크게 다르지 않다. 그러나 계산비용을 고려한다면 최대가능도 추정이 선호될 수 있다. 한편, 베이지안 추정에서 회귀계수들의 사후분포는 표본크기와는 별 관계없이 정규분포에 매우 근사하였으며 표본크기가 클 경우 오차항 분산의 사후분포도 로그변환을 통해 정규성을 확보할 수 있었다. 여기서 문제가 되는 부분은 표본크기가 작은 경우 각 추정방법의 표준오차의 적절한지에 대한 평가가 필요하다. 이를 위해 4절에서는 표본크가가 상대적으로 작을 때 각 추정방법의 신뢰구간에 대한 포함확률을 모의실험을 통해 추정하기로 한다. 표준오차값이 적절한 값이려면 신뢰구간의 포함확률이 명목신뢰수준에 근사하여야 할 것이다.

4. 신뢰구간에 대한 모의실험과 결론

표준오차의 적절성 평가를 위해

$$y = \max\{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon, 0\}$$

과 같은 토빗모형에서 표본크기가 각각 $n = 30, 50, 100, 200$인 데이터를 발생시키고 최대가능도 추정, 준결례사전분포 및 무정보 사전분포에 의한 베이지안 추정을 실시하였다. 이러한 과정을 2000회 반복하여 추정값 및 표준오차의 평균, 신뢰구간들의 포함확률의 추정값을 구하였으며 이를 Table 4.1에 수록하였다. 이때 $\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, \log\sigma)$는 $(1, 1, 1, 0)$으로 주어졌으며, x_1과 x_2는 각각 -1과 1의 값이 확률 1/2인 베르누이 난수와 표준정규난수로부터 생성하였다. 또, 오차항 ϵ도 표준정규난수로부터 생성되었다. 이러한 데이터의 발생은 Bilias 등 (2000)과 Yu와 Stander (2007)의 모의실험에서도 사용된 것으로 절단된 데이터의 비율이 약 30% 정도이다.

모의실험 결과 표본크기에 관계없이 최대가능도 추정의 평균값은 베이지안 추정의 평균값은 실제 모수값에 매우 근접하고 있으며 각 추정방법에 따른 추정값의 차이도 유의하지 않은 것으로 나타났다. 또, 표본크기가 커짐에 따라 값의 차이는 더욱 감소하는 것을 볼 수 있다. 추정값의 평균이 실제모수에 가깝다는 것은 추정량의 편의가 작다는 것을 의미하므로 빈도학파에서 선호되는 성질이다. 그러나, 이 결과가 최대가능도 추정값이 실제 값에 가깝다는 의미는 아니다. 베이지안 추정량은 편의추정량이기 때문에 추정값의 평균을 구할 경우 최대가능도 추정보다 실제값에서 멀어질 수 있다. Chib (1992)의 모의 실험에서는 베이지안 추정값이 최대가능도 추정보다 실제값에 가깝다고 하였기 때문에 여기에서는 추정 자

Table 4.1. Averages of estimate, coverage and width of confidence interval with relatively small sample size.

Method	$\boldsymbol{\theta}$	Estimate	SE	Wald(HPD) Coverage	Wald(HPD) Width	Posterior Coverage	Posterior Width	Estimate	SE	Wald(HPD) Coverage	Wald(HPD) Width	Posterior Coverage	Posterior Width
		$n = 30$						$n = 50$					
MLE	β_0	0.994	0.196	0.922	0.769			0.993	0.156	0.931	0.611		
	β_1	0.997	0.192	0.924	0.754			0.996	0.152	0.935	0.595		
	β_2	0.997	0.189	0.919	0.742			1.002	0.153	0.931	0.598		
	$\log\sigma$	−0.094	0.163	0.896	0.639			−0.050	0.125	0.919	0.489		
Semi-conjugate	β_0	0.988	0.233	0.953	0.914	0.937	0.921	0.994	0.170	0.948	0.668	0.935	0.671
	β_1	1.035	0.227	0.950	0.889	0.945	0.897	1.015	0.165	0.953	0.649	0.951	0.652
	β_2	1.042	0.223	0.951	0.876	0.944	0.883	1.024	0.166	0.946	0.653	0.942	0.656
	$\log\sigma$	0.022	0.183	0.956	0.717	0.952	0.716	0.013	0.133	0.956	0.522	0.950	0.522
Non-informative	β_0	0.991	0.235	0.958	0.922	0.945	0.929	0.996	0.172	0.952	0.676	0.946	0.680
	β_1	1.037	0.229	0.962	0.896	0.956	0.905	1.026	0.167	0.952	0.655	0.944	0.659
	β_2	1.056	0.225	0.965	0.883	0.960	0.891	1.030	0.168	0.952	0.659	0.947	0.663
	$\log\sigma$	0.032	0.183	0.960	0.717	0.952	0.717	0.021	0.133	0.949	0.522	0.948	0.522
		$n = 100$						$n = 200$					
MLE	β_0	0.999	0.111	0.934	0.436			0.999	0.080	0.948	0.313		
	β_1	1.000	0.109	0.944	0.428			1.002	0.079	0.948	0.311		
	β_2	1.006	0.104	0.943	0.408			1.003	0.082	0.947	0.323		
	$\log\sigma$	−0.022	0.086	0.931	0.335			−0.009	0.060	0.939	0.233		
Semi-conjugate	β_0	0.987	0.116	0.941	0.453	0.937	0.454	0.993	0.081	0.950	0.319	0.951	0.319
	β_1	1.008	0.114	0.952	0.445	0.950	0.446	1.006	0.081	0.952	0.316	0.952	0.317
	β_2	1.014	0.108	0.948	0.424	0.946	0.425	1.007	0.084	0.948	0.329	0.950	0.329
	$\log\sigma$	0.006	0.088	0.954	0.346	0.951	0.346	0.004	0.061	0.948	0.237	0.947	0.237
Non-informative	β_0	0.984	0.116	0.951	0.454	0.952	0.455	0.994	0.081	0.952	0.319	0.949	0.319
	β_1	1.014	0.114	0.957	0.445	0.955	0.447	1.006	0.081	0.947	0.317	0.948	0.317
	β_2	1.011	0.108	0.957	0.424	0.955	0.425	1.006	0.084	0.950	0.329	0.948	0.329
	$\log\sigma$	0.005	0.088	0.953	0.345	0.953	0.345	0.004	0.060	0.954	0.237	0.951	0.237

$\boldsymbol{\theta} = (1, 1, 1, 0)$

체 보다는 표준오차의 적절성 평가를 위해 포함확률과 명목신뢰구준과의 차이에 대해 살펴보기로 한다.

이 모의실험에서도 평균적으로 최대가능도 추정의 표준오차는 베이지안 추정의 표준오차보다 작았다. 이에 따라 최대가능도 신뢰구간의 넓이는 베이지안 신뢰구간보다 짧았으며, 2000회의 반복에서 추정된 포함확률이 베이지안 신뢰구간보다 작았으며 포함확률이 명목신뢰수준에 근사하다고 평가하기 어렵다. 특히 표본크기가 작은 $n = 30$의 경우, 포함확률이 명목신뢰수준인 0.95보다 현저하게 작은 것을 확인할 수 있다. 표에서 표본크기가 커짐에 따라 근사성이 개선되어 가는 것을 볼 수 있으나 $n = 200$인 경우에도 포함확률이 명목신뢰수준보다 작은 경향이 있어 표준오차가 근소하게 과소 추정된 것으로 판단된다. 이에 반해 베이지안 신뢰구간의 포함확률은 대체적으로 명목신뢰수준인 0.95에 근사하기 때문에 표본크기가 작은 경우도 베이지안 추정의 표준오차는 적절한 것으로 평가할 수 있다. 그러므로 베이지안 신뢰구간이 최대가능도 추정보다도 빈도학파적 효율성이 높다고 할 수 있다

표준오차와 포함확률의 근접성에 대한 베이지안 방법간의 비교에서는 유의미한 차이를 발견하지 못하였다. 즉, 모의실험에서 사용된 준결레 사전분포는 무정보사전분포와 매우 유사하기 때문에 추정에서 큰 차이가 없었다. 또, 사후분포의 백분위점과 HPD와 비교에서도 유의미한 차이가 있다고 보기 어려우나 계산의 간편성을 고려하면 HPD 방법이 선호된다고 하겠다.

References

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

Bilias, Y., Chen, S. and Ying, Z. (2000). Simple resampling methods for censored regression quantiles, *Journal of Econometrics,* **68**, 303–338.

Chib, S. (1992). Bayesian inference in the Tobit censored regression model, *Journal of Econometrics,* **51**, 77–99.

Green, W. H. (1990). *Econometric Analysis*, Macmillan, New York.

Henningsen, A. (2012). *Estimating Censored Regression Models in R using the censReg Packages*, http://cran.r-project.org/package=censReg.

Lee, S.-C. and Lee, D. (2002). Bayesian analysis of multivariate threshold animal models using Gibbs sampling, *Journal of the Korean Statistical Society,* **31**, 177–198.

Lee, S.-C. (2006). Interval estimation of binomial proportions based on weighted Polya posterior, *Computational Statistics & Data Analysis,* **51**, 1012–1021.

Mroz, T. A. (1987). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions, *Econometrica,* **55**, 765–799. </PRE></BODY></HTML>

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities, *Journal of American Statistical Association,* **81**, 82–86.

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

Yu, K. and Stander, J. (2007). Bayesian analysis of a Tobit quantile regression model, *Journal of Econometrics,* **137**, 260–276.

Wooldridge, J. M. (2009) *Introductory Econometrics: A Modern Approach*, South-Werstern Cengage Learning.

토빗회귀모형에서 베이지안 구간추정

이승천[a,1] · 최병수[b]

[a]한신대학교 응용통계학과, [b]한성대학교 멀티미디어학과

(2013년 7월 22일 접수, 2013년 10월 21일 수정, 2013년 10월 21일 채택)

요 약

Tobin (1958)에 의해 처음 소개된 절단 회귀모형에서 베이지안 추정은 최대가능도 추정보다 실제값에 가까운 것으로 알려져 있으나 베이지안 방법론이 구간추정 문제에 있어서도 성공적으로 작동할 수 있을 지에 대해서는 알려진 바가 없다. 일반적으로 베이지안 방법론에서 사전분포는 분석자의 사전정보를 반영하기 때문에 주관적인 분석이 될 수 밖에 없는데, 이렇게 주관적인 분석에서는 빈도학파들이 요구하는 기준을 따르기 어렵다. 그러나 무정보사전분포는 때때로 빈도학파적 특성을 갖는 베이지안 추론을 가능하게 한다. 본 연구에서는 절단 회귀모형에서 무정보사전분포에 의한 베이지안 신뢰구간의 빈도학파적 특성을 살펴보고 최대가능도 추정 신뢰구간과 포함확률을 비교한다. 이를 통해 최대가능도 추정의 표준오차가 과소 추정되고 있음 밝힌다.

주요용어: 깁스샘플링, 무정보사전분포, 절단회귀모형, 포함확률.

이 논문은 한신대학교 학술연구비 지원에 의하여 연구되었음.
이 논문은 한성대학교 학술연구비 지원에 의하여 연구되었음.
[1]교신저자: (447-791) 경기도 오산시 양산동 411, 한신대학교 응용통계학과, 교수. E-mail: seung@kss.or.kr