응용통계연구 제10권 1호
1997년 3월 pp. 151-161

몬테칼로깁스표본기법을 이용한 누적로짓 모형의 베이지안 분석[1)]

오 만 숙[2)]

요 약

순서적 다항자료의 누적로짓 모형에 대한 베이지안 사후추론을 위하여 몬테칼로 깁스표본기법을 제안하였다. 원래의 모형에서는 깁스표본기법 적용에 필수적으로 요구되는 각 원소모수의 조건부 확률분포가 난수생성에 편리한 형태로 주어지지 않으므로 Albert and Chib(1993)과 Oh(1997)에서 이항 로짓모형에 사용한 바와 같이 적절한 잠재변수를 도입하여 깁스표본기법 적용에 매우 편리한 형태를 갖도록 한다.

1. 서론

많은 실제적인 문제에서 자료가 이항이나 순서가 있는 다항의 형태로 주어진다. 이항자료에 대한 가장 보편적인 모형은 로짓모형이며 순서적 다항자료에 대한 모형은 누적로짓 모형이 널리 쓰이고 있는데 이는 이들 모형의 적합성이 뛰어나고 의미있는 해석이 가능하기 때문이다.

이들 모형에 대한 베이지안 분석은 중요한 사전정보가 존재하거나 표본크기가 크지 않은 경우, 모수에 제한조건이 있는 경우에 매우 유용하게 적용될 수 있다. 그러나 베이지안 분석의 적용에 있어서 실제적인 문제는 사후분포의 형태가 간단치 않아 필요한 사후정보의 도출에 필요한 적분의 계산이 용이하지 않다는 것이다. 이러한 문제의 해결책에 대한 연구가 꾸준히 이루어지고 있는 바 Zeller and Rossi(1984)는 로짓모형에 대한 베이지안 분석을 위하여 몬테칼로 주표본기법을 사용하였으며 Zeger and Karim(1991), Albert and Chib(1993), Dellaportas and Smith(1993), Oh(1997)는 이항자료에 대한 베이지안 분석을 위하여 몬테칼로 깁스표본기법을 제안하였다. 이 중 특히 Albert and Chib(1993)과 Oh(1997)는 주어진 모형에 적절한 잠재변수를 도입하여 전체 모형을 깁스표본기법의 적용이 용이한 단순모형으로 만드는 방법을 제시하였다. 이같이 잠재변수를 도입하는 방법의 장점은 모형을 연속 잠재변수에 대한 정규선형 모형으로 간주할 수 있어 새로운 모형에 대한 깁스표본기법의 적용이 누구나 사용할 수 있게 간단하다는 것이다.

그러나 이상의 연구들은 Albert and Chib(1993)을 제외하고는 모두 이항자료에 대한 것으로 다항자료에의 직접적인 적용은 불가능하였다. 그리고 Albert and Chib(1993)에서도 특수한 형태의 다항모형만 언급되었을 뿐 보편적으로 널리 쓰이는 누적로짓 모형에 적용할 수 있는 것은

1) 이 논문은 1995 학술진흥재단 자유공모 연구비에 의하여 이루어진 것임.
2) (120-750) 서울시 서대문구 대현동 이화여자대학교 통계학과 조교수

아니었다.

본 논문에서는 다항자료의 누적로짓 모형에 쉽게 적용할 수 있는 깁스표본기법을 제안하고자 한다. 기본 아이디어는 Albert and Chib(1993)과 Oh(1997)에서와 같이 적절한 연속 잠재변수를 도입하여 깁스표본기법의 적용이 용이한 새로운 모형을 찾는 것이다.

본 논문의 2절에서는 누적로짓 모형에 대하여 기술하고 3절에서는 도입될 잠재변수에 대하여 소개하고 새로운 모형에서 각 모수에 대한 조건부 분포를 유도하며 4절에서는 제안된 기법의 적용예가 주어진다. 마지막으로 맺음말이 5절에 주어진다.

2. 누적로짓 모형

순서적 다항자료가 r개의 행과 c개의 열을 가진 $r\times c$ 돗수분포표에 정리되어 있다고 가정하자. 돗수분포표에서 i번째 행과 j번째 열에 해당하는 항 (i,j)의 확률을 π_{ij}라 하고 (i,j) 항에 속하는 자료의 돗수를 n_{ij}라 정의한다. 자료의 수집에서 독립다항모형을 가정하면 각 행 i에 대하여

$$\sum_{j=1}^{c}\pi_{ij} = 1$$

이 성립한다.

위와 같은 순서적 다항자료에 대한 누적로짓 모형은 각 항의 확률 π_{ij}와 p차원의 모수 $\boldsymbol{\beta}$ 사이에 다음과 같은 관계식을 가정한다:

$$\begin{aligned}
\pi_{i1} &= 1-F(\boldsymbol{x}_{i1}^{t}\boldsymbol{\beta})\\
\pi_{i1}+\pi_{i2} &= 1-F(\boldsymbol{x}_{i2}^{t}\boldsymbol{\beta})\\
&\vdots\\
\pi_{i1}+\ldots.+\pi_{i,c-1} &= 1-F(\boldsymbol{x}_{i,c-1}^{t}\boldsymbol{\beta}).
\end{aligned}\qquad (1)$$

위 식에서 F는 로지스틱 분포의 누적분포함수이며 $\boldsymbol{x}_{ij}$ 는 (i,j) 항에 대응하는 p개의 독립변수의 값으로 이루어진 p차원 벡터이다. 식 (1)로부터

$$F(\boldsymbol{x}_{i1}^{t}\boldsymbol{\beta}) > F(\boldsymbol{x}_{i2}^{t}\boldsymbol{\beta}) > \ldots. > F(\boldsymbol{x}_{i,c-1}^{t}\boldsymbol{\beta})$$

가 성립함을 알 수 있고 또한 누적분포함수의 성질로부터

$$\boldsymbol{x}_{i1}^{t}\boldsymbol{\beta} > \boldsymbol{x}_{i2}^{t}\boldsymbol{\beta} > \ldots. > \boldsymbol{x}_{i,c-1}^{t}\boldsymbol{\beta}$$

임을 알 수 있다. 누적로짓 모형에서 $c=2$ 인 경우가 바로 이항자료에 대한 로짓 모형이다.

누적로짓 모형에 대한 베이지안 분석은 미지의 모수 $\boldsymbol{\beta}$의 사후분포, 특히 $\boldsymbol{\beta}$의 각 원소의

사후 주변분포에 기초한다. β의 사후밀도함수는 β의 사전밀도함수를 $\pi(\beta)$라 하고 $F(x_{i0}^t \beta)=1$, $F(x_{ic}^t \beta)=0$ 라 정의할 때 다음과 같이 주어진다 :

$$\pi(\beta|Data) \propto \pi(\beta) \prod_{i=1}^{r} \prod_{j=1}^{c} \pi_{ij}^{n_{ij}}$$

$$= \pi(\beta) \prod_{i=1}^{r} \prod_{j=1}^{c} (F(x_{i,j-1}^t \beta) - F(x_{ij}^t \beta))^{n_{ij}}. \quad (2)$$

그런데 베이지안 사후추론의 난점은 식 (2)에서 보는 바와 같이 β의 사후밀도함수가 간단한 형태가 아니어서 필요한 사후추정량의 계산에 필요한 적분의 계산이 수리적으로 불가능하고 따라서 수치적 적분기법의 도입이 필수적으로 요구된다는 점이다.

통계에서 요구되는 적분의 계산에 적용될 수 있는 많은 수치적분 기법 중 몬테칼로 적분기법은 최근에 들어서 그 적용의 용이성과 계산속도의 향상으로 주목을 받고 있는데 그 중 특히 깁스표본기법은 복잡한 다차원분포로부터의 난수생성에 매우 유용한 기법으로 알려져 있다. 깁스표본기법 알고리즘은 Gelfand and Smith (1990, 1992)에 자세히 기술되어 있다.

그러나 누적로짓 모형의 경우 깁스표본기법의 직접적인 적용이 불가능한데 이는 식 (2)로부터 알 수 있는 바와 같이 β의 각 원소 모수의 조건부 분포가 난수생성이 쉬운 형태로 주어지지 않기 때문이다. 이러한 문제를 해결하기 위하여 본 논문에서는 적절한 잠재변수를 도입하여 각 원소 모수의 조건부 분포를 난수생성이 용이한 형태로 만들고자 한다.

3. 잠재변수의 도입과 조건부 분포

먼저 로지스틱 분포에 대한 t분포의 근사를 고려하기로 하자. Albert and Chib(1993)에서 언급된 바와 같이 로지스틱 분포는 자유도 8, 위치모수 0, 척도모수 $1/0.634^2$을 갖는 t분포, $t_8(0, 1/0.634^2)$, 로 매우 정확히 근사시킬 수 있다 (그림 1 참조). 따라서 로지스틱분포를 $t_8(0, 1/0.634^2)$ 분포로 간주하여 Albert and Chib(1993) 에서와 유사한 형태의 잠재변수를 고려할 수 있다. 로지스틱 분포를 t분포로 근사시킬 때의 커다란 장점은 t분포를 정규분포와 역감마분포의 합성으로 나타낼 수 있기 때문에 아래에서 보는 바와 같이 β의 우도함수가 정규사전분포와 짝관계로서 쉽게 결합될 수 있다는 점이다.

그러나 누적로짓 모형에서의 난점은 각 행 마다 c($c \geq 2$)개의 범주 혹은 열이 존재한다는 것이다. 즉, 각 행 마다 관측치가 속할 수 있는 범주가 c개이므로 잠재변수의 구간을 c개로 나누어 각 관측치가 어느 범주에 속하는지에 따라 대응하는 잠재변수가 c개의 구간 중 하나에 속하도록 결정되어져야 한다는 것이다. 이는 구간을 0을 중심으로 두 개의 구간으로 나누는 이항로짓 모형에서와 달리 잠재변수의 구간이 모수 β에 의존하도록 하는 결과를 낳는다. 두번째

의 난점은 이항자료에서는 각 행 마다 하나의 독립변수의 조합이 주어졌으나 다항자료에서는 각 행 마다 $c-1$개의 조합이 주어진다는 것이다. 즉, 행 i에서 $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,c-1}$가 $\boldsymbol{\beta}$와 연관되어 있으므로 행 i에 대응하는 잠재변수의 위치모수로서 $\boldsymbol{x}_{i1}^t\boldsymbol{\beta}, \ldots, \boldsymbol{x}_{i,c-1}^t\boldsymbol{\beta}$ 를 모두 고려해야 한다.

이상의 점들을 고려하여 다음과 같은 잠재변수를 제안한다. 먼저, 행 i에 속하는 각 범주의 돗수 n_{ik}, $k=1,\ldots,c$, 를 대략 균등하게 $c-1$개로 나누어 이를 n_{ikj}, $j=1,\ldots,c-1$, 라 한다. 예를 들면, $[a]$를 a의 정수부분을 나타낸다고 할 때

$$n_{ikj}=[\frac{n_{ik}}{c-1}], \quad j=1,\ldots,c-2$$

$$n_{ikj}=n_{ik}-(c-2)[\frac{n_{ik}}{c-1}], \quad j=c-1$$

로 놓을 수 있겠다. 다음, $1\le i\le r$, $1\le k\le c$, $1\le j\le c-1$, $1\le l\le n_{ikj}$인 각 i, j, k, l에 대하여 잠재변수 z_{ijkl}와 λ_{ijkl}을 다음과 같이 정의한다.

$$z_{ijkl}|\lambda_{ijkl},\ \boldsymbol{\beta} \sim N(\boldsymbol{x}_{ij}^t\boldsymbol{\beta}, \frac{1}{\lambda_{ijkl}0.634^2}) \cdot I(-\boldsymbol{x}_{i,k-1}^t\boldsymbol{\beta}<z_{ijkl}-\boldsymbol{x}_{ij}^t\boldsymbol{\beta}<-\boldsymbol{x}_{ik}^t\boldsymbol{\beta}), \tag{3}$$

$$\lambda_{ijkl} \sim \Gamma(8/2,\ 2/8). \tag{4}$$

위 식에서 $N(\mu,\sigma^2)$는 평균 μ, 분산 σ^2을 갖는 정규분포, $\Gamma(\delta,\gamma)$는 모수 δ,γ를 갖는 감마분포를 나타내고 $I(\cdot)$은 지시함수이며 $\boldsymbol{x}_{i0}^t\boldsymbol{\beta}=\infty$, $\boldsymbol{x}_{ic}^t\boldsymbol{\beta}=-\infty$로 정의된다. 이는 t분포가 정규분포와 역감마분포의 합성으로 나타낼 수 있는 점을 이용한 것으로 (3)과 (4)로부터 z_{ijkl}의 주변분포가 로지스틱 분포의 근사인 $t_8(0,1/0.634^2)$이 되도록 한 것이다. 따라서 $\boldsymbol{\beta}$가 주어졌을 때 식 (3)에서 주어진 제한조건의 조건부 확률을 구하면, $T_{8,0.634}(\cdot)$을 $t_8(0,\frac{1}{0.634^2})$의 누적분포함수라 할 때,

$$\begin{aligned}
&P(-\boldsymbol{x}_{i,k-1}^t\boldsymbol{\beta} < z_{ijkl}-\boldsymbol{x}_{ij}^t\boldsymbol{\beta} < -\boldsymbol{x}_{ik}^t\boldsymbol{\beta}|\ \boldsymbol{\beta})\\
&= T_{8,0.634}(-\boldsymbol{x}_{ik}^t\boldsymbol{\beta}) - T_{8,0.634}(-\boldsymbol{x}_{i,k-1}^t\boldsymbol{\beta})\\
&= 1-T_{8,0.634}(\boldsymbol{x}_{ik}^t\boldsymbol{\beta}) - [1-T_{8,0.634}(\boldsymbol{x}_{i,k-1}^t\boldsymbol{\beta})]\\
&\approx 1-F(\boldsymbol{x}_{ik}^t\boldsymbol{\beta}) - [1-F(\boldsymbol{x}_{i,k-1}^t\boldsymbol{\beta})]\\
&= \pi_{ik}
\end{aligned}$$

로, z_{ijkl}는 위치모수 $x_{ij}^t\beta$를 갖고 확률 π_{ik}를 갖는 구간에 속하며 이러한 z_{ijkl}의 수는 n_{ik}에 비례함을 알 수 있다. 다시 말하면, 항 (i,k)에 속하는 관측치에 대응하는 연속 잠재변수는 모두 n_{ik} 개로서 이들은 위치모수로 $x_{ij}^t\beta$, $j=1,..,c-1$, 를 가지며 이들이 속하는 구간의 확률은 항 (i,k)의 확률과 일치하도록 한 것이다.

구체적으로 식 (3)에서의 z_{ijkl}의 도입을 $c=3$이고 n_{ik}모두 짝수인 경우에 예를 들면, $n_{ik1}=n_{ik2}=n_{ik}/2$ 이므로,

$$
\begin{aligned}
z_{i11l}|\lambda_{i11l},\beta &\sim N(x_{i1}^t\beta, \frac{1}{\lambda_{i11}0.634^2})\, I(z_{i11l}<0), \quad l=1,\ldots,n_{i1/2}\\
z_{i12l}|\lambda_{i12l},\beta &\sim N(x_{i1}^t\beta, \frac{1}{\lambda_{i12}0.634^2})\, I(0<z_{i12l}<-x_{i2}^t\beta+x_{i1}^t\beta), \quad l=1,\ldots,n_{i2/2}\\
z_{i13l}|\lambda_{i13l},\beta &\sim N(x_{i1}^t\beta, \frac{1}{\lambda_{i13}0.634^2})\, I(-x_{i2}^t\beta+x_{i1}^t\beta<z_{i13l}), \quad l=1,\ldots,n_{i3/2}\\
z_{i21l}|\lambda_{i21l},\beta &\sim N(x_{i2}^t\beta, \frac{1}{\lambda_{i21}0.634^2})\, I(z_{i21l}<-x_{i1}^t\beta+x_{i2}^t\beta), \quad l=1,\ldots,n_{i1/2}\\
z_{i22l}|\lambda_{i22l},\beta &\sim N(x_{i2}^t\beta, \frac{1}{\lambda_{i22}0.634^2})\, I(-x_{i1}^t\beta+x_{i2}^t\beta<z_{i22l}<0), \quad l=1,\ldots,n_{i2/2}\\
z_{i23l}|\lambda_{i23l},\beta &\sim N(x_{i2}^t\beta, \frac{1}{\lambda_{i23}0.634^2})\, I(0<z_{i23l}), \quad l=1,\ldots,n_{i3/2} \qquad (5)
\end{aligned}
$$

이다. 즉, $n_{i.}/2=\sum_{k=1}^{c} n_{ik}/2$개의 잠재변수는 $x_{i1}^t\beta$를 위치모수로 갖고 나머지 $n_{i.}/2$개의 잠재변수는 $x_{i2}\beta$를 위치모수로 가지며 각 $n_{i.}/2$개의 잠재변수가 π_{i1}, π_{i2}, π_{i3}의 확률을 갖는 세 개의 구간 중 하나에 속하게 되는데 π_{ik}확률을 갖는 구간에 속하는 잠재변수의 수가 관측돗수 n_{ik}와 일치하도록 한 것이다.

식 (3)과 (4)의 잠재변수의 도입은 연속 잠재변수 z_{ijkl}이 모수 β와 이산 관측치 n_{ik}를 연결하는 역할을 하여 모형을 마치 연속변수 z_{ijkl}를 관측치로 하는 정규 선형모형 처럼 만든다. 따라서 새로운 모형에서 z_{ijkl}, λ_{ijkl}, β를 모수로 간주할 때 각 모수의 조건부 분포가 아래에서 보는 바와 같이 난수생성에 매우 쉬운 형태로 주어진다.

식 (3)과 (4)로부터 z_{ijkl}, λ_{ijkl}, β의 조건부 분포를 구하면 λ_{ijkl}과 β가 주어졌을 때 z_{ijkl}의 조건부 분포는 식 (3)과 같고, z_{ijkl}과 β가 주어졌을 때 λ_{ijkl}의 조건부 분포는

$$\lambda_{ijkl}|z_{ijkl},\ \beta \ \sim\ \Gamma((8+1)/2,\ 2/(8+0.634^2(z_{ijkl}-\boldsymbol{x}_{ij}^t\beta)^2)) \tag{6}$$

이다. 그리고 β의 사전분포로 $N(\beta_0, A)$를 가정하면,

$$z=(\boldsymbol{z}_{111}^t,\ldots,\boldsymbol{z}_{r,c-1,c}^t)^t,\quad \boldsymbol{z}_{ijk}=(z_{ijk1},\ldots,z_{ijkn_{ijk}})^t,\quad W=diag(\lambda_{ijkl})$$

가 주어졌을 때 β의 조건부 분포는

$$\beta|\ z,\ \lambda\ \sim\ N(\ (X^tWX+A^{-1})^{-1}(X^tWz+A^{-1}\beta_0),\ (X^tWX+A^{-1})^{-1}\ )$$

$$\cdot I(\prod_{i,j,k}[-\boldsymbol{x}_{i,k-1}^t\beta+\boldsymbol{x}_{ij}^t\beta<\min_l z_{ijkl}<\max_l z_{ijkl}<-\boldsymbol{x}_{ik}^t\beta+\boldsymbol{x}_{ij}^t\beta]) \tag{7}$$

로 주어진다. 식 (7)에서 $n\times p$, $n=\sum_{i,k}n_{ik}$, 차원 행렬 X는

$$X=(\ 1_{n_{1,1}}\otimes\boldsymbol{x}_{11}^t,\ldots,1_{n_{1,c-1}}\otimes\boldsymbol{x}_{1,c-1}^t,\ldots..,1_{n_{r,1}}\otimes\boldsymbol{x}_{r1}^t,\ldots,1_{n_{r,c-1}}\otimes\boldsymbol{x}_{r,c-1}^t)$$

이며 여기에서 1_m은 m개의 1로 구성된 열벡터이고 $n_{i,j=}\sum_k n_{ikj}$ 이며 $\otimes$는 Kronecker곱을 나타낸다. 특히 β에 대하여 무정보 사전분포를 가정하면 A를 포함한 항들이 없어지며 제한조건을 제외한 정규분포의 평균과 공분산 행렬이 각각 $(X^tWX)^{-1}X^tWz$, $(X^tWX)^{-1}$로 바뀌게 된다.

식 (3), (6) 에서 보면 z_{ijkl}과 λ_{ijkl}의 난수생성은 매우 간단하다. 반면 β의 조건부 분포 (7)는 제한이 있는 다차원 정규분포로 이는 이항로짓 모형의 경우와 다른 특징이다. 그러나 이 제한조건도 $k=1$, $k=c$, $k=j$, 또는 $k=j+1$인 경우에는 식 (5)에서 보는 바와 같이 매우 단순화되며 그렇지 않은 경우라도 β에 대한 선형 제한조건이므로 β의 각 원소에 대한 조건부 분포를 쉽게 유도해낼 수 있다. 예를 들어 $c=3$인 경우 β의 제한조건은 식 (5)로부터

$$\max\{\max_l z_{i1l},-\min_l z_{i2l}\}\ <\ \boldsymbol{x}_{i1}^t\beta-\boldsymbol{x}_{i2}^t\beta\ <\ \min\{\min_l z_{i3l},-\max_l z_{i2l}\}$$

로 매우 간단히 요약됨을 알 수 있다.

이상에서 살펴본 바와 같이 제안된 잠재변수의 도입으로 새로운 모수 z_{ijkl}, λ_{ijkl}, β의 조건부 분포들이 모두 난수생성이 쉬운 형태로 주어지므로 식 (3), (6), (7)의 차례로 난수를 생성하는 깁스표본기법을 손쉽게 적용할 수 있으며 이로부터 β의 원소변수들의 주변사후분포를 추정하여 베이지안 사후추론에 사용할 수 있다.

4. 적용 예

Goodman(1968)과 Agresti(1984)는 417명의 위궤양 환자들을 네 그룹으로 나누어 각 그룹 마다 서로 다른 수술을 실시한 후 수술 후유증의 하나인 dumping의 정도에 따라 분류한 돗수분포표를 보여준다. 본 논문에서는 제안한 잠재변수의 도입과 깁스표본기법의 적용 예를 들기 위하여 원래의 돗수분포표에 나타난 돗수를 대략 5로 나눈 아래의 돗수분포표를 사용하기로 한다.

이 자료에 대한 모형으로 Agresti(1984)가 제안한 누적로짓에 대한 균일 결합모형

$$\log\left[\frac{\pi_{i,j+1}+\cdots+\pi_{ic}}{\pi_{i1}+\cdots\pi_{ij}}\right]=\beta_j+\beta_3(i-2.5),\quad i=1,..,4,\ j=1,2$$

을 사용하기로 한다. 즉, $\boldsymbol{\beta}=(\beta_1,\beta_2,\beta_3)$ 이고 $\boldsymbol{x}_{i1}^t=(1,\ 0,\ i-2.5)$, $\boldsymbol{x}_{i2}^t=(0,\ 1,\ i-2.5)$ 이며 $\pi_{i1}+\cdots\pi_{ij}=1-F(\boldsymbol{x}_{ij}^t\boldsymbol{\beta})$ 의 관계가 성립된다.

표 1: 위궤양 수술과 후유증에 대한 돗수분포표

	Dumping severity			
Operation	None	Slight	Moderate	Total
A	12	6	2	20
B	14	4	2	20
C	12	8	2	22
D	10	8	4	22
	48	26	10	84

위 모형에 제안된 깁스표본기법을 적용하기 위하여 먼저 n_{ik1}, n_{ik2}를 구하면 $c=3$이고 n_{ik}가 모두 짝수이므로 각 n_{ik}를 균등하게 양분하여 $n_{ik1}=n_{ik2}=n_{ik}/2$로 둘 수 있다. 다음, 식 (3)에서 주어진 z_{ijkl}의 조건부 분포로부터 난수를 생성하기 위하여 λ_{ijkl}의 초기치를 모두 1로 두고 $\boldsymbol{\beta}$의 초기치로는 $\boldsymbol{\beta}$의 최우추정치인 $(-0.3, -2.1, 0.26)$을 사용한다. z_{ijkl}의 난수생성 후 식 (6)의 λ_{ijkl}의 조건부 분포로부터 난수를 생성하고, 다음으로 $\boldsymbol{\beta}$의 난수를 식 (7)의 조건

부 분포로부터 생성한다. 이 과정을 계속적으로 반복하면서 매 100회 째의 β난수들로부터 Q-Q(Quantile-Quantile) Plot을 그려본 결과 1000회에서 Q-Q Plot이 거의 기울기 1의 직선 형태로 나타나므로 1000회에서 대략적인 깁스표본기법의 수렴이 일어나는 것으로 간주할 수 있다. 따라서 1000회 이후 10000개의 난수를 더 생성하여 이들로부터 β_1, β_2, β_3의 주변확률밀도함수를 추정하였으며 그 결과를 그림 2, 3, 4에 표시하였다. 그림 2-4에서 실선은 β_i의 실제 주변확률밀도함수로 충분히 큰 숫자의 표본을 갖는 몬테칼로 주표본기법을 적용시켜 얻은 것이다. 모든 경우에 프로그램 언어로는 FORTRAN 5.1을 사용하였고 컴퓨터 기종으로는 IBM PC 486 DX2를 사용하였다.

그림에서 보면 제안된 깁스표본기법이 β_i의 주변밀도함수들을 매우 근접하게 추정하고 있음을 알 수 있다. 특히 β_1과 β_3의 경우는 사용된 10000개의 난수보다 상당히 작은 숫자의 난수로서도 대략 같은 정도의 정확도를 갖는 추정치를 얻을 수 있었다. 이로 미루어 β_2의 수렴속도가 β_1과 β_3에 비하여 느림을 짐작할 수 있다. 이는 β_2와 β_1, β_3의 상관계수가 각각 0.2265, -0.1531로서 β_1과 β_3의 상관계수인 -0.0899에 비하여 상대적으로 큰데 기인하지 않나 추측된다.

또한 깁스표본기법으로부터 얻은 난수들은 실제 사후분포로부터 생성된 난수들로 간주될 수 있기 때문에 주변밀도함수 이외의 다른 사후추정량들, 예를 들면 사후평균, 사후분산 등, 을 이들 난수를 사용하여 손쉽게 추정할 수 있을 것이다.

5. 맺음말

본 논문에서는 순서적 다항자료의 누적로짓 모형에 대한 베이지안 분석에 필요한 수치적분을 위해 잠재변수 도입을 통한 깁스표본기법을 제안하였다. 이는 Albert and Chib(1993)과 Oh(1997)의 이항자료에 대한 깁스표본기법을 확장한 것으로 특히 로지스틱 분포에 대한 t분포의 근사를 이용하여 매우 간단한 잠재변수로서 문제를 해결하였다. 잠재변수의 도입으로 필요한 조건부 분포들이 모두 기존의 알고리즘을 이용한 난수생성에 용이한 형태를 갖게 되어 깁스표본기법의 적용이 누구나 사용하기 쉽다는 것이 제안된 기법의 핵심이다.

로지스틱 분포에 대한 좀 더 정확한 근사를 위해서는 Oh(1997)에서 사용한 부분정규근사를 사용할 수도 있겠으나 이는 로짓모형 이외의 모형에도 사용할 수 있는 일반성이 있는 반면 t분포의 근사에 비하여 복잡한 면이 있고 또한 본 논문에서 다루는 누적로짓 모형에 대해서는 t분포가 만족할 만한 근사를 제공하기 때문에 제안된 바와 같은 t근사로 충분하다고 생각된다.

참 고 문 헌

[1] Agresti, A. (1984). *Analysis of Ordinal Categorical Data,* John Wiley & Sons, New York.

[2] Albert, J. and Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data: A Gibbs Sampling Approach, *Journal of the American Statistical Association*, Vol. 88, 669-679.

[3] Dellaportas, P. and Smith, A.F.M. (1993). Bayesian Inference for Generalized Linear and Proportional Hazards Models via Gibbs Sampling, *Applied Statistics*, Vol. 42, 443-459.

[4] Devroye, L. (1986). *Non-Uniform Random Variate Generation*, New York : Springer-Verlag.

[5] Gelfand, A. and Smith, A.F.M. (1990). Sampling Based Approaches to Calculating Marginal Densities, *Journal of the American Statistical Association*, Vol. 85, 398-409.

[6] Gelfand, A. and Smith, A.F.M. (1992). Bayesian Analysis of Constrained Parameter and Truncated Data Problems Using Gibbs Sampling, *Journal of the American Statistical Association*, Vol. 87, 523-532.

[7] Goodman, L.A. (1968). The Analysis of Cross-Classified Data: Independence, Quasi Independence, and Interaction in Contingency Tables with or without Missing Cells, *Journal of the American Statistical Association*, Vol. 63, 1091-1131.

[8] McCullagh, P. and Nelder, J.A. (1988). *Generalized Linear Models,* New York: Chapman and Hall.

[9] Oh, M-S. (1997), A Gibbs Sampling Approach to Bayesian Analysis of Generalized Linear Models for Binary Data, *Computational Statistics*, to appear.

[10] Zeger, S.L. and Karim, M.R. (1991). Generalized Linear Models with Random Effects: a Gibbs Sampling Approach, *Journal of the American Statistics,* Vol. 86, 79-86.

[11] Zeller, A. and Rossi, P.E. (1984). Bayesian Analysis of Dichotomous Quantal Response Data, *Journal of Econometrics*, Vol. 25, 365-393.

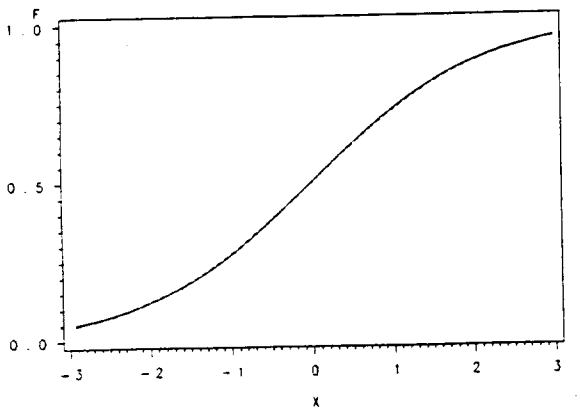

그림 1: $t_8(0, 1/0.634^2)$분포(- - - -)와
로지스틱분포(———)의 누적분포함수

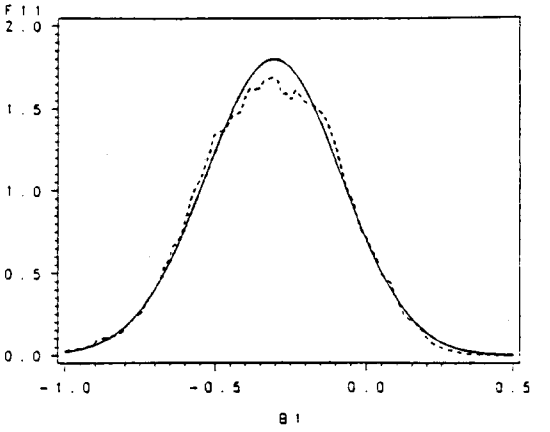

그림 2: β_1의 주변밀도함수:
실제(———), 추정(- - - -)

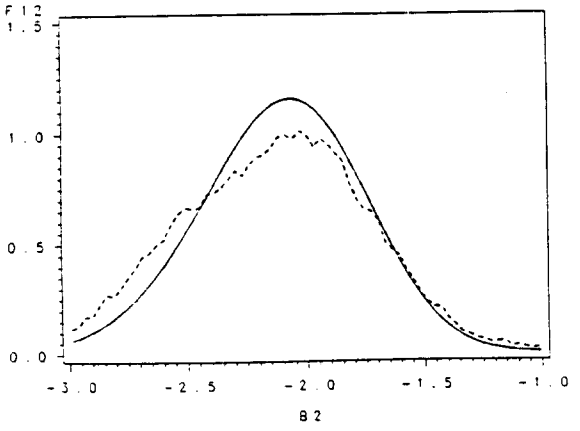

그림 3: β_2의 주변밀도함수:
실제(———), 추정(- - - -)

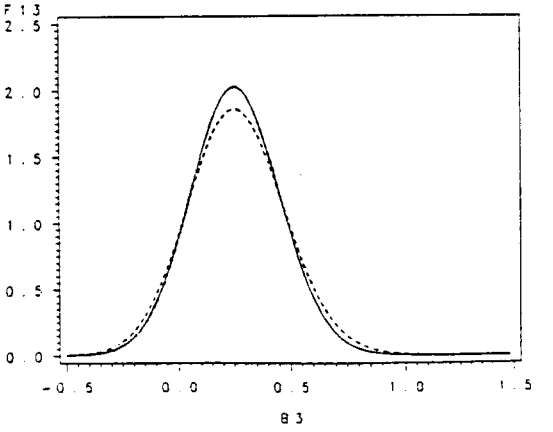

그림 4: β_3의 주변밀도함수:
실제(———), 추정(- - - -)

Bayesian Analysis of Cumulative Logit Models Using the Monte Carlo Gibbs Sampling[3)]

Man-Suk Oh[4)]

Abstract

An easy Monte Carlo Gibbs sampling approach is suggested for Bayesian analysis of cumulative logit models for ordinal polytomous data. Because in the cumulative logit model the posterior conditional distributions of parameters are not given in convenient forms for random sample generation, appropriate latent variables are introduced into the model so that in the new model all the conditional distributions are given in very convenient forms for implementation of the Gibbs sampler.

3) This research was supported by Non Directed Research Fund, Korea Research Foundation,

4) Assistant Professor, Dept. of Statistics, Ewha Womans University, So-Dae-Mun Gu, Seoul (120-750).