응용통계연구 제 7권 1호
1994년 3월 pp. 1 - 17

중도 절단 자료에서의 역추정 문제[1)]

박래현,[2)] 이석훈,[3)] 이낙영,[4)] 박영옥,[5)] 이상호[6)]

요약

본 논문은 종속변수가 중도 절단된 경우 역추정을 베이지안 방법으로 접근하였는데 이때 특별히 Gibbs Sampling을 응용하여 사후분포를 계산하는 것을 토의하였다. 적용의 예로서 실제 자료에서의 점추정 및 Simulated Annealing을 이용한 구간추정도 하였다.

1. 서론

회귀모형에서 중요하게 다루는 예측 중 하나는 주어진 독립변수 x_0에 대해 종속변수 y_0 또는 $E(y_0)$를 추론하는 것이다. Calibration(역추정)문제는 역으로 y_0가 주어졌을 때 미지의 x_0을 점추정 또는 구간 추정하는 예측 문제로써 통계적 bioassay, probit분석 등과 밀접한 관계가 있고 화학, 의학, 공학등에 두루 적용되어 왔다.

x_0를 추론하기 위해 여러 학자들이 연구를 해 왔는데 고전적인 non-Bayesian 방법과 베이지안(Bayesian) 방법의 두 종류로 나눌 수 있다. 고전적인 방법들은 신뢰구간도 어떤 조건하에서만 구해지는 반면 베이지안 방법은 고전적 방법의 단점을 극복한 것이라 할 수 있다.

Calibration문제에서 우리들이 고민해온 점은 종속변수가 중도 절단된(Censored) 경우를 생각할 수 있는 데 이를 해결할 일반적인 방법이 제안되어 있지 않다는 것이다. 회귀분석에서 자료가 중도 절단된 경우 Miller(1976), Buckley와 James(1979)등이 생각해 낸 Non-Bayesian 방법과 Wei와 Tanner(1990)가 제안한 베이지안 방법들을 발견할 수 있다. Wei와 Tanner의 방법은 회귀계수의 사후분포를 Tanner와 Wong(1987)에 의해서 제안된 Data Augmentation 알고리즘이란 컴퓨터 시뮬레이션 방법을 통해 계산한 점이 핵심이 된다. 우리는 중도 절단된 자료를 가진 Calibration모형에서 x_0를 점추정 또는 구간추정을 수행할 때 Non-Bayesian방법과 Data Augmentation 알고리즘이 역추정이라는 특수성 때문에 해결하기가 어려운 부분을 Gibbs sampling 방법을 이용하여 해결하는 과정을 보였다.

Gibbs sampling 방법은 Geman과 Geman(1984)에 의해 소개되어 컴퓨터 과학분야에 널리 쓰여왔고 통계학에서는 Gelfand와 Smith(1990)가 처음 도입하였으며 특히 베이지안 적분에 넓은 적용 범위를 갖고 있어 베이지안 적분 계산을 위한 다른 방법들인 Tierney와 Kadane(1986) 또

1) 이 논문은 1991년도 교육부 지원 한국학술진흥재단의 대학부설연구소 지원 학술연구 조성비에 의하여 연구되었음.
2) (305-764) 대전시 유성구 궁동 220 충남대학교 통계학과
3) (305-764) 대전시 유성구 궁동 220 충남대학교 통계학과
4) (305-764) 대전시 유성구 궁동 220 충남대학교 통계학과
5) (305-764) 대전시 유성구 궁동 220 충남대학교 통계학과 박사과정
6) (305-764) 대전시 유성구 궁동 220 충남대학교 통계학과 박사과정

는 Naylor와 Smith(1982)등의 근사기법(Approximation methods), Data Augmentation 알고리즘, Rubin(1987)의 Importance Sampling 등으로는 해결하기 힘든 것을 풀어 줄 수 있는데 그 한 예가 우리의 연구 목표인 x_0에 대한 사후분포 p(x_0|censored data)의 계산이라 생각된다.

제 2절에서는 본 논문의 목적인 사후분포 p(x_0|censored data)를 추정하기 위해 우리가 사용한 방법인 Gibbs sampling에 대한 소개와 함께 이 방법의 몇 가지 특성 및 성질에 대해 언급하였고 제 3절에서는 중도 절단된 자료를 포함하지 않는 역추정 문제에서 x_0의 사후분포를 계산하기 위한 Gibbs Sampling 방법과 이에 관련된 필요한 이론들을 다루었다. 제 4절에서는 본 연구의 핵심인 중도 절단된 자료를 가진 경우 x_0의 사후분포를 계산하기 위한 Gibbs Sampling을 자세히 다루었고 제 5절에서는 중도 절단된 것이 없는 자료와 있는 자료를 가지고 x_0의 사후분포를 추정하였으며 Simulated Annealing을 이용하여 각각의 경우 구간추정도 하였다.

2. Gibbs Sampling

우리의 최종 목표인 중도 절단된 자료를 가진 단순 선형 Calibration모형하에서 x_0의 사후분포 p(x_0|censored data)를 계산하기 위해 사용한 방법인 Gibbs sampling에 대해 알아보기로 하겠다.

$U_1, U_2, \ldots, U_k$를 우리가 다루고 있는 k개의 확률변수라 하고 FCD(full conditional distribution)라 불리우는 k개의 분포 $p(U_1|U_2, \ldots, U_k), \ldots, p(U_2|U_1, U_3, \ldots, U_k), \ldots, p(U_k|U_1, \ldots, U_{k-1})$가 우리에게 이용가능(available)하다고 가정하자. 여기서 "이용가능"의 의미는 각 FCD $p(U_r|U_1, \ldots, U_{r-1}, U_{r+1}, \ldots, U_k)$, $r = 1, \ldots, k$로부터 조건변수(conditioning variables)의 값이 주어졌을 경우 U_r의 표본들을 더 이상의 조작없이 효율적으로 얻을 수 있다 라는 뜻이다. Gibbs Sampling Algorithm 이란 위와 같은 모든 FCD들로 부터 U_r의 주변분포 p(U_r)의 추정을 하는 일종의 Monte Carlo 방법인데 Geman과 Geman(1984)에 의해 자세히 소개되었고, 신경회로망(neural networks), 이미지 재생(image reconstruction), 전문가 시스템(expert system)등 많은 변수를 가지고 주어진 어떤 일을 처리해야만 하는 컴퓨터 분야에 널리 적용되어 왔다. Gibbs Sampling을 통계학 분야의 계산에 처음 소개한 사람은 Gelfand와 Smith(1990)로 여러 사후분포를 계산하는 예의 소개와 함께, Gibbs Sampling과 다른 두 Monte Carlo방법인 Tanner와 Wong(1987)의 Data Augmentation algorithm, Rubin(1987)의 Importance Sampling사이의 관계를 자세히 다루었다.

Gibbs Sampling은 다음과 같은 Markovian updating scheme이다.

$U_1^{(0)}, U_2^{(0)}, \ldots, U_k^{(0)}$을 $U_1, U_2, \ldots, U_k$의 초기치라 할 때 첫번째 iteration을 아래와 같이 한다.

$p(U_1 \mid U_2^{(0)}, U_3^{(0)}, \ldots, U_k^{(0)})$을 따르는 하나의 관찰값 $U_1^{(1)}$을 생성

$p(U_2 \mid U_1^{(1)}, U_3^{(0)}, \ldots, U_k^{(0)})$을 따르는 하나의 관찰값 $U_2^{(1)}$을 생성

$p(U_3 \mid U_1^{(1)}, U_2^{(1)}, U_4^{(0)}, \ldots, U_k^{(0)})$을 따르는 하나의 관찰값 $U_3^{(1)}$을 생성

$\vdots$

$p(U_k \mid U_1^{(1)}, U_2^{(1)}, \ldots, U_{k-1}^{(1)})$을 따르는 하나의 관찰값 $U_k^{(1)}$을 생성

그리고 $U_1^{(i)}, \ldots, U_k^{(i)}$를 생성하기 위한 i번째 iteration은 $U_1^{(i-1)}, \ldots, U_k^{(i-1)}$을 초기치로 하는 위와 같은 scheme이다. Geman과 Geman(1984)은 $i \to \infty$일때 $(U_1^{(i)}, \ldots, U_k^{(i)})$가 $(U_1, \ldots, U_k)$로 분포수렴(converge in distribution), 즉 $U_r^{(i)}$가 U_r로 분포수렴함을 증명하였다.

상기와 같은 첫번째부터 i-th iteration까지를 m번 반복하여 데이타 $(U_{1j}^{(i)}, \ldots, U_{kj}^{(i)})$, $j = 1, \ldots, m$을 얻으면 (이 경우 총 mik개의 random variates를 생성해야함) 우리의 목적이라 할 수 있는 Ur의 주변분포 $p(U_r)$은 다음처럼 추정할 수 있다.

$$\hat{p}(U_r) = \frac{1}{m} \sum_{j=1}^{m} p(U_r \mid U_{1j}^{(i)}, U_{2j}^{(i)}, \ldots, U_{(r-1)j}^{(i)}, U_{(r+1)j}^{(i)}, \ldots, U_{kj}^{(i)})$$

Gibbs Sampling 방법의 장점은 개념상 단순하고 implementation이 용이할 뿐아니라 모든 FCD를 이용가능한 문제에는 쉽게 적용가능 하므로 적용범위가 매우 넓은 점이라 할 수 있겠다. 실제로 우리가 여기서 다루려 하는 중도 절단된 자료에서의 역추정문제는 모든 FCD를 쉽게 이용할 수 있어 Gibbs Sampling 방법으로 편하게 해결할 수 있었지만 다른 조건부 분포인 $p(U_1, \ldots, U_{i-1}, U_{i+1}, \ldots, U_k \mid U_i)$, i=1,...,k 가 이용가능하지 않아 Data Augmentation Algorithm의 적용이 불가능했고, Importance Sampling도 성공적인 적용 방법을 찾을 수 없었다. Gibbs Sampling 방법을 적용한 논문들로는 Zeger와 Karim(1991), Gelfand et al (1990)등을 들 수 있다.

Gibbs Sampling Method를 적용할 때 조사되어야 할 내용중 하나는 어떤 조건하에서 모든 FCD들이 결합분포를 유일하게 결정하느냐 하는 것인데, 이에 대한 해답을 Geman과 Geman(1984)이 자세히 다루었다. 우리가 다루고 있는 $p(x_0 \mid \text{censored data})$의 추정을 포함한 많은 통계문제들은 결합 사후분포가 유일하게 정의되어 있고 이 결합 사후분포로 부터 모든 이용가능한 FCD들을 계산해 낼 수 있으므로 상기와 같은 문제는 해결된 것이라 할 수 있다.

3. 중도절단이 안된 자료에서의 역추정

우리가 여기서 다루려고 하는 Calibration모형은 다음과 같다.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

$$y_{0j} = \beta_0 + \beta_1 x_0 + \varepsilon_{0j}, \quad j = 1, \ldots, k \qquad (3.1)$$

위에서 흔히 하듯이 x_i, x_0는 실험자가 조절가능한 수학적 변수이고 yi, y_{0j}는 확률변수로 가정하겠으며 오차항 ε_i, ε_{0j} ($i=1,\ldots,n$, $j=1,\ldots,k$)들은 IID(Independent and Identically Distributed)인 $N(0,\sigma^2)$의 분포를 따르는 확률변수로 하겠다. 앞에서 언급했듯이 Calibration문제란 주어진 y_{0j} , j=1,…,k에 대해 미지의 x_0를 점추정 혹은 구간추정하는 것인데 접근 방법으로는 크게 Non-Bayesian 방법과 Bayesian 방법이 있다. Non-Bayesian 방법의 큰 단점은 Hoadley(1970), Park 과 Lee(1990)등이 언급했듯이 구간추정시 폐쇄형 구간(closed interval)이 나오지 않는 경우가 있으며, 폐쇄형 구간을 구할 수 있는 경우에도 신뢰수준을 줄 수 없다는 것이다. 이에 반해 베이지안 접근은 사후분포 $p(x_0|\text{data})$를 (여기서 data란 (x_i , y_i), $i=1,\ldots,n$, y_{0j} , $j=1,\ldots,k$ 를 의미) 바탕으로 구간추정을 하는 것이므로 x_0의 정확한 신뢰구간을 어느 경우에나 구할 수 있지만 Hoadley(1970)와 Brown(1982)에서 알 수 있듯이 비정보적 사전분포에서는 $p(x_0|\text{data})$가 improper posterior가 되므로 x_0의 사전분포를 정보적인 것으로 줄 수 밖에 없다는 점이다. Hunter와 Lamboy(1981)의 베이지안 접근도 저자 자신들은 비정보적 임을 주장하지만, Hill(1981)에서 알 수 있듯이 결국은 정보적인 사전분포를 x_0에 준 것이라 할 수 있다.

본 절의 목적은 비정보적인 일양 사전분포(Uniform Prior)와 정보적인 정규분포의 사전분포를 x_0에 주었을 경우 각각 사후분포 $p(x_0|\text{data})$를 Gibbs Sampling 방법으로 추정하고 이를 바탕으로 x_0를 점추정, 구간추정하는 것이다. 비정보적인 사전분포하에서 Hoadley(1970)가 구한 사후분포가 x_0의 사후분포이지만 improper 사후분포이어서 이것으로는 x_0를 구간추정할 수 없기 때문에 대안으로 Gibbs Sampling을 이용하여 사후분포를 추정한 것을 가지고 x_0를 구간추정하려한다.

3.1 일양 사전분포의 경우

우리가 다루는 문제에 관련된 모수들은 x_0, β_0, β_1, σ^2인데 이들에 대한 prior를 Hoadley(1970)에서처럼 $p(x_0,\beta_0,\beta_1,\sigma^2) = p(x_0)\,p(\beta_0,\beta_1,\sigma^2) \propto 1/\sigma^2$로 하려한다. 즉 x_0와 (β_0, β_1, σ^2)은 독립이며 x_0와 (β_0, β_1, σ^2)에 대한 prior를 비정보적인 것으로 하겠다.이 경우 (x_0, β_0, β_1, σ^2)의 사후분포는

$$p(x_0,\beta_0,\beta_1,\sigma^2 \mid \text{data}) \propto p(x_0,\beta_0,\beta_1,\sigma^2)\, p(\underline{y},\underline{y_0}, \mid x_0,\beta_0,\beta_1,\sigma^2)$$

$$\propto \sigma^{-2} p(\underline{y} \mid \beta_0,\beta_1,\sigma^2)\, p(\underline{y_0}, \mid x_0,\beta_0,\beta_1,\sigma^2)$$

으로 된다. 위에서 $\underline{y} = (y_1,\ldots,y_n)'$, $\underline{y_0} = (y_{01},\ldots,y_{0k})'$ 이고 data란 ($\underline{y}$, $\underline{y_0}$))를 의미한다. 이 경우 우리가 필요로 하는 3개의 FCD는 아래처럼 구해진다.

(1) $p(x_0 \mid \beta_0,\beta_1,\sigma^2,\text{data})$;

$$p(x_0 \mid \beta_0,\beta_1,\sigma^2,\text{data}) \propto p(x_0 \mid \beta_0,\beta_1,\sigma^2,\underline{y})\; p(\underline{y_0} \mid \beta_0,\beta_1,\sigma^2,\underline{y},x_0)$$

$$= p(x_0)\; p(\underline{y_0} \mid \beta_0,\beta_1,\sigma^2,x_0)$$

$$\propto p(x_0)\; p(\overline{y_0} \mid \beta_0,\beta_1,\sigma^2,x_0)$$

$$\propto \exp\left[-\frac{k}{2\sigma^2}(\overline{y_0}-\beta_0-\beta_1 x_0)^2\right]$$

$$\propto \exp\left[-\frac{k\beta_1^2}{2\sigma^2}\left\{x_0-\left(\frac{\overline{y_0}-\beta_0}{\beta_1}\right)\right\}^2\right] \qquad \text{if} \quad \beta_1 \neq 0$$

위에서 첫번째 등식이 성립하는 이유는 Hoadley(1970)가 언급했듯이Calibration에서 β_0, β_1, σ^2, $\underline{y}$는 Calibration체계에 관련된 것이고 x_0는 현재 알고 싶어하는 대상물(object)에 관련된 것이므로 이들은 서로 독립이라 가정할 수 있기 때문이고 두번째 비례식이 성립하는 이유는 β_0, β_1, σ^2이 주어진 경우는 $\overline{y_0} = k^{-1}\sum_{j=1}^{k} y_{0j}$가 x_0의 충분통계량이기 때문이다.

윗식에서 볼 때 결국 $p(x_0 \mid \beta_0,\beta_1,\sigma^2,\text{data})$는 $\beta_1 \neq 0$이면 $N[(\overline{y_0}-\beta_0)/\beta_1\ ,\ \sigma^2/k\beta_1^2]$ 밀도 함수라는 것을 알았다.

(2) $p(\sigma^2 \mid \beta_0,\beta_1,x_0,\underline{y},\underline{y_0})$;

$$s^{*2} = \frac{1}{n+k}\left[\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2+\sum_{i=1}^{k}(y_{0j}-\beta_0-\beta_1 x_0)^2\right]$$

라고 하면 β_0, β_1, x_0가 주어진 후는 s^{*2}이 σ^2의 충분통계량이 되므로

$$p(\sigma^2 \mid \beta_0,\beta_1,x_0,\underline{y},\underline{y_0}) \propto p(\sigma^2 \mid s^{*2})$$

$$\propto p(\sigma^2)\; p(s^{*2} \mid \sigma^2)$$

$$\propto \sigma^{-2}\; p(s^{*2} \mid \sigma^2)$$

위에서 $(n+k)s^{*2}/\sigma^2 \sim \chi^2(n+k)$ 로 부터 우리는 다음을 얻을 수 있다. 여기서 $\chi^2(n+k)$란 자유도가 (n+k)인 카이자승분포를 뜻한다.

$$p(s^{*2} \mid \sigma^2) \propto (\sigma^{-2})\left(\frac{(n+k)s^{*2}}{\sigma^2}\right)^{((n+k)/2)-1} \cdot \exp\left[-\frac{1}{2}\left(\frac{(n+k)s^{*2}}{\sigma^2}\right)\right]$$

따라서 우리는 다음을 얻는다.

$$p(\sigma^2 \mid \beta_0,\beta_1,x_0,\underline{y},\underline{y_0}) \propto (\sigma^2)^{-((n+k)/2)-1} \cdot \exp\left[-\frac{1}{2}\left(\frac{(n+k)s^{*2}}{\sigma^2}\right)\right] \quad (3.2)$$

실제 문제에서는 우리는 (3.2)식을 이용해서 σ^2의 random variate를 생성하기보다는 $(n+k)s^{*2}/\sigma^2$ 의 분포가 $\chi^2(n+k)$분포이므로 자유도 (n+k)의 χ^2분포를 따르는 random variate u를 생성한후 $\sigma^2 = (n+k)s^{*2}/u$로써 σ^2을 생성하는 방법을 택하려 한다.

(3) $p(\beta_0,\beta_1 \mid x_0,\sigma^2,\underline{y},\underline{y_0})$

앞에서도 언급했듯이 β_0, β_1은 Calibration체계에서 나온 것이고 $\underline{y_0}$, x_0는 현재 Calibrate 하려는 대상물에 관련된 것이므로 이들은 서로 독립이라 할 수 있고 β_0, β_1의 최소자승 추정치인

$\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = (X'X)^{-1}X'\underline{y}$ 가 β_0, β_1의 충분통계량이므로 다음과 같이 구할 수 있다.

$$\begin{aligned} p(\beta_0,\beta_1 \mid x_0,\sigma^2,\underline{y},\underline{y_0}) &= p(\beta_0,\beta_1 \mid ,\sigma^2,\underline{y}) \\ &= p(\beta_0,\beta_1 \mid ,\sigma^2,b_0,b_1) \\ &\propto p(\beta_0,\beta_1)\, p(b_0,b_1 \mid ,\beta_0,\beta_1,\sigma^2) \end{aligned}$$

위에서 β_0, β_1, σ^2이 주어졌을 경우 b_0 , b_1은 이변량 정규분포 $N_2((\beta_0,\beta_1)'$, $(X'X)^{-1}\sigma^2)$를 따르므로 결국 $p(\beta_0,\beta_1 \mid x_0,\sigma^2,\underline{y},\underline{y_0})$는 $N_2((b_0,b_1)'$, $(X'X)^{-1}\sigma^2)$의 밀도함수가 됨을 알 수 있다. 여기서 $X' = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix}$ 이다.

3.2 정규분포의 사전분포 경우

앞에서 토의한 것처럼 x_0의 사전분포를 비정보적인 improper 일양사전분포로 잡으면 결국 improper 사후분포가 얻어지어 x_0의 구간추정을 실질적으로 어렵게 만든다. 그래서 결국은 x_0의 사전분포를 정보적으로 할 수 밖에 없는데 우리는 x_0의 사전분포를 $N(\mu,\tau^2)$ (μ , τ^2은 주어진 값)으로 하고 나머지 β_0, β_1, σ^2의 prior는 앞에서 잡은 비정보적인 사전분포를 주기로 하겠다. 또 전과 같이 x_0와 (β_0, β_1, σ^2)은 서로 독립인 것으로 가정하자. 이 경우 다른 두

FCD는 전과 같고 다만 $p(x_0 \mid \beta_0,\beta_1,\sigma^2,\underline{y},\underline{y_0})$만 N(a,b)의 밀도함수로 전의 것과 다르다는 것을 쉽게 알 수 있다.

위에서

$$a = \left[\frac{k\beta_1^2}{\sigma^2}+\frac{1}{\tau^2}\right]^{-1}\left[\frac{k\beta_1^2}{\sigma^2}\left\{\frac{\overline{y_0}-\beta_0}{\beta_1}\right\}+\frac{1}{\tau^2}\mu\right]$$

$$b = \left[\frac{k\beta_1^2}{\sigma^2}+\frac{1}{\tau^2}\right]^{-1}$$

이다.

3.3 $p(x_0 \mid \text{data})$의 추정

$(\beta_0, \beta_1, \sigma^2, x_0)$의 초기치를 $(\beta_0^{(0)},\beta_1^{(0)},x_0^{(0)},\sigma^{2(0)})$으로 잡고 앞에서의 FCD들을 이용해서 i번째까지의 iteration을 m번씩 반복하여 다음의 생성값들을 얻었다고 하자.

$$(\beta_{0j}^{(i)},\ \beta_{1j}^{(i)},\ x_{0j}^{(i)},\ \sigma_j^{2(i)}),\ j=1,\ldots,m$$

상기 생성값을 이용하여 우리의 목적인 두가지 종류의 사전분포하에서 $p(x_0|\text{data})$를 아래와 같이 추정하면 된다.

$$\hat{p}(x_0 \mid data) = \frac{1}{m}\sum_{j=1}^{m} p(x_0 \mid \beta_{0j}^{(i)},\beta_{1j}^{(i)},\sigma_j^{2(i)})$$

위에서 $p(x_0 \mid \beta_{0j}^{(i)},\beta_{1j}^{(i)},\sigma_j^{2(i)})$는 앞에서 선택한 두 종류의 사전분포에 따라 $N[(\overline{y_0}-\beta_{0j}^{(i)})/\beta_{1j}^{(i)},\ \sigma_j^{2(i)}/k[\beta_{1j}^{(i)}]^2]$또는 $N(a_j^{(i)},\ b_j^{(i)})$의 밀도함수이다.

여기서

$$a_j^{(i)} = \left[\frac{k(\beta_{1j}^{(i)})^2}{\sigma_j^{2(i)}}+\frac{1}{\tau^2}\right]^{-1}\left[\frac{k(\beta_{1j}^{(i)})^2}{\sigma_j^{2(i)}}\left\{\frac{\overline{y_0}-\beta_{0j}^{(i)}}{\beta_{1j}^{(i)}}\right\}+\frac{1}{\tau^2}\mu\right]$$

$$b_j^{(i)} = \left[\frac{k(\beta_{1j}^{(i)})^2}{\sigma_j^{2(i)}}+\frac{1}{\tau^2}\right]^{-1}$$

이다.

4. 중도절단 자료에서의 역추정

본절에서 우리가 도입하는 모형은 다음과 같다.

$$T_i = \beta_0+\beta_1 x_i+\varepsilon_i,\ i=1,\ldots,n$$

$$T_{0j} = \beta_0+\beta_1 x_0+\varepsilon_{0j}\ , j=1,\ldots,k \qquad (4.1)$$

위에서 종속변수는 right-censored로 되어 있는, 즉 c_1 , $c_2,\ldots,$ c_n , $c_{01},\ldots,$ c_{0k}를 censoring time이라 할 때

$$y_i = \min(T_i\ ,\ c_i)\ ,\ i=1,\ldots,n$$

$$y_{0j} = \min(T_{0j}\ ,\ c_{0j})\ ,\ j=1,\ldots,k$$

인 경우를 다루려고 하는데 사전분포는 앞 절의 두 종류를 그대로 쓰려한다. 여기서 censoring은 type I censoring이건 random censoring이건 우리의 접근 방법이 자료는 주어진 것으로 보는 베이지안 방법이므로 다음에서 보듯이 우리의 이론 전개에는 전혀 영향을 주지 않는다. 표기를 간편하게 하기 위해 $\underline{T} = (T_1,\ldots,T_n)'$, $\underline{T_0} = (T_{01},\ldots,T_{0k})'$ 로 표시하기로 하고 data인 (y_1,δ_1) ,...,(y_n,δ_n),(y_{01},δ_{01}) ,...,(y_{0k},δ_{0k})를 "censored data"로 표현하자. 여기서

$$\delta_i = \begin{cases} 1\ , & T_i \le c_i\text{인 경우(실제값 } T_i\text{가 중도절단이 안된 경우)} \\ 0\ , & T_i > c_i\text{인 경우(실제값 } T_i\text{가 중도절단이 된 경우)} \end{cases} \quad i=1,\ldots,n$$

$$\delta_{0j} = \begin{cases} 1\ , & T_{0j} \le c_{0j}\text{인 경우} \\ 0\ , & T_{0j} > c_{0j}\text{인 경우} \end{cases} \quad j=1,\ldots,k$$

이다.

p(x_0|censored data)를 추정하기 위해 우리가 도입한 접근의 기본 생각은 Wei와 Tanner(1990)에 바탕을 둔 것으로 y_i 혹은 y_{0j}가 중도절단된 경우 실제값인 T_i , T_{0j}를 미지의 모수로 놓고 (β_0, β_1), x_0, σ^2의 FCD뿐만 아니라 T_i, T_{0j}의 FCD도 구하여 이들로 부터 전 절과 같이 Gibbs Sampling 방법의 iteration으로 생성값을 구해서 목적한 사후분포를 추정하는 것이다. 필요한 5개의 FCD중 (β_0, β_1), x_0, σ^2에 대한 것은 $\underline{y}$, $\underline{y_0}$, $\overline{y_0}$,를 $\underline{T}$, $\underline{T_0}$, $\overline{T_0} = k^{-1}\sum_{j=1}^{k} T_{0j}$으로 대치시키는 것 이외에 는 전 절의 결과와 똑같음을 쉽게 알 수 있다. 중도 절단된 경우 T_i, T_{0j}를 생성하기위한 FCD는 전 절에서 도입한 두 종류의 사전분포에는 관계없이 다음과 같이 유도된다.

T_i와 T_{0j}는 서로 독립이므로 p(T_i | β_0, β_1, x_0, σ^2,censored data)와 p(T_{0j} | β_0, β_1, x_0, σ^2,censored data)를 구하면 충분한데 censored data중 T_i, T_{0j}의 분포에 영향을 주는 사실은 $T_i > c_i$, $T_{0j} > c_{0j}$라는 것 뿐이므로 결국 p(T_i | β_0, β_1, x_0, σ^2, $T_i > c_i$)와 p(T_{0j} | β_0, β_1, x_0, σ^2, $T_{0j} > c_{0j}$)가 우리가 필요로 하는 FCD가 된다. 먼저 T_i의 FCD를 T_i의 조건부 누적분포를 통해서 다음처럼 구해보기로 한다.

$$
\begin{aligned}
p(T_i \le t \mid \beta_0,\beta_1,x_0,\sigma^2,T_i > c_i) \ (\text{단 } t > c_i \text{ 임}) \\
&= p(T_i \le t \mid \beta_0,\beta_1,\sigma^2,T_i > c_i) \\
&= \frac{p(c_i < T_i \le t \mid \beta_0,\beta_1,\sigma^2)}{p(T_i > c_i \mid \beta_0,\beta_1,\sigma^2)} \\
&= \frac{p(w_i < Z_i \le z \mid \beta_0,\beta_1,\sigma^2)}{p(Z > w_i \mid \beta_0,\beta_1,\sigma^2)} \\
&= \frac{\Phi(z)-\Phi(w_i)}{1-\Phi(w_i)} \ (\text{단 } z > w_i \text{ 임})
\end{aligned}
\tag{4.2}
$$

위에서

$$Z_i = \sigma^{-1}(T_i-\beta_0-\beta_1 x_i)$$

$$w_i = \sigma^{-1}(c_i-\beta_0-\beta_1 x_i)$$

$$z_i = \sigma^{-1}(t_i-\beta_0-\beta_1 x_i) \text{ 이고}$$

$\Phi(x), \phi(x)$는 각각 표준정규분포의 누적분포함수와 밀도함수를 표시한다. 위 (4.2)를 z에 대해 미분하면 Z_i의 FCD인 $\phi(z)/(1-\Phi(w_i))$ (단 $z > w_i$)를 얻는데 Z_i의 FCD를 통해서 T_i를 생성하면 된다. 물론 T_i절단된 경우가 아니면 $T_i = y_i$이므로 T_i를 생성할 필요는 없다.

마찬가지 방법으로 하면 중도절단된 T_{0j}를 생성하기 위해서는 먼저 $\phi(s)/(1-\Phi(w_{0j}))$ (단 $s > w_{0j}$)에서 Z_{0j}를 생성한 후 $T_{0j} = \sigma Z_{0j}+\beta_0+\beta_1 x_0$를 통해 생성하면 된다. 여기서 $w_{0j} = \sigma^{-1}(c_{0j}-\beta_0-\beta_1 x_0)$이며 중도절단이 아닌 경우는 $T_{0j} = y_{0j}$로 하면 된다.

$(\beta_0^{(0)},\beta_1^{(0)},x_0^{(0)},\sigma^{2(0)},\underline{T}^{(0)},\underline{T_0}^{(0)})$를 초기치라 하고 5개의 FCD를 이용해서 i번째까지의 iteration을 m번씩 반복하여 다음의 생성값을 얻었다 하자.

$$(\underline{T_j}^{(i)},\underline{T_{0j}}^{(i)},\beta_{0j}^{(i)},\beta_{1j}^{(i)},x_{0j}^{(i)},\sigma_j^{2(i)}) \quad j = 1,\ldots,m$$

위의 생성값을 바탕으로 p(x_0|censored data)를 아래와 같이 추정하면 될 것이다.

$$\hat{p}(x_0 \mid \textit{censored data}) = \frac{1}{m}\sum_{j=1}^{m} p(x_0 \mid \beta_{0j}^{(i)},\beta_{1j}^{(i)},x_{0j}^{(i)},\sigma_j^{2(i)},\underline{T_{0j}}^{(i)},\underline{T_j}^{(i)})$$

여기에서 $p(x_0 \mid \beta_{0j}^{(i)},\beta_{1j}^{(i)},x_{0j}^{(i)},\sigma_j^{2(i)},\underline{T_{0j}}^{(i)},\underline{T_j}^{(i)})$는 선택한 두 종류의 사전분포에 따라

$N\left[\frac{\overline{T_{0j}^{(i)}}-\beta_{0j}^{(i)}}{\beta_{1j}^{(i)}} , \frac{\sigma_j^{2(i)}}{k[\beta_{1j}^{(i)}]^2}\right]$의 밀도함수, 또는 $N(a_j^{*(i)} , b_j^{(i)})$의 밀도함수이다.

위에서 $\overline{T_{0j}^{(i)}} = k^{-1}\sum_{r=1}^{k} T_{0rj}^{(i)}$ 이고 $T_{0rj}^{(i)}$는 $\underline{T_{0j}^{(i)}}$의 r번째 원소이며 $a_j^{*(i)}$는 $\overline{y_0}$대신 $\overline{T_{0j}^{(i)}}$를 대치시킨 3.3절의 $a_j^{(i)}$값이다.

5. 적 용 예

5.1 중도절단된 자료가 아닌 경우

우리가 분석해보려는 자료는 Lwin과 Maritz(1980)에 있는 자료로 토양에 포함된 수분의 비율(%)을 잴 때 비용과 시간은 많이 들지만 매우 정확한 방법인 실험실 방법으로 잰 값(x)과 덜 정확하지만 시간과 비용이 매우 적게드는 On site 방법으로 잰 값(y)으로 다음 표 5.1에 나와있다.

표 5.1

실험실 방법(x)	On Site 방법	실험실 방법(x)	On Site 방법
23.7	35.3	24.3	33.1
20.2	27.6	10.6	12.8
24.5	36.2	15.2	23.1
15.8	21.6	11.4	19.6
29.2	39.8	19.7	26.1
17.8	24.1	12.7	19.3
10.1	16.1	31.8	39.8
19.0	27.5		

위 자료를 바탕으로 어떤 견본 토양에서 On Site 방법으로 잰 수분의 비율이 y_0=12.6(k=1의 경우임)일 때 진짜 수분의 비율(x_0)의 사후분포를 추정해보려 하는데 모형은 3절을 따르고 x_0에 대한 사전분포는 비정보적인 일양분포, $N(0,10^2)$과 $N(0,100^2)$의 세 종류를 주어 구하려 한다. 세번째 사전분포에서 분산을 크게 준 이유는 비정보적에 가까우면서도 비정보적인 일양분포의 사전분포하에서는 improper 사후분포가 나오는 문제점을 제거할 수 있기 때문이다. 세 종류의 사전분포하에서 x_0의 사후분포를 Gibbs Sampling 방법으로 추정한 것이 그림 5.1인데 iteration 수 i와 반복수 m은 더 늘려도 결과에 변동이 없으므로 i=30, m=50으로 하였다. y_0=12.6에 대한 x_0의 실제 값은 18.8인데 그림 5.1을 보면 일양사전분포와 $N(0,100^2)$의 사전분포하에서 나온 두 사후분포는 모두 18.8근처에서 비교적 밀집되어 분포되어 있고 두 분포가 겹쳐서 구별이 거의 힘든 것을 알 수 있으며 $N(0,10^2)$의 사전분포하에서 구한 사후분포는 사전

분포의 평균인 0의 영향을 받아서 다른 두 분포에 비해 왼쪽에 위치해 있다. 또 세개의 사후분포를 바탕으로 x_0를 점추정(일반화 MLE로 추정했음)과 구간추정을 한 결과 (구간추정의 결과는 Simulated Annealing 방법으로 구한 95% HPD(Highest Posterior Density) 신뢰구간임)와 고전적인 Non-Bayesian 방법으로 구한 점추정및 구간추정 (점추정은 $\widehat{x_0} = b_1^{-1}(y_0 - b_0)$로 했고 구간추정은 Fieller(1932) 방법을 이용한 95% 신뢰구간임)의 결과가 표 5.2에 있다.

표 5.2

방법	점추정	95% 신뢰구간
고전적 방법	17.35	(12.87 , 21.99)
베이지안 방법 (일양분포의 사전분포)	18.35	(13.44 , 22.74)
베이지안 방법 ($N(0,10^2)$의 사전분포)	17.70	(12.63 , 21.78)
베이지안 방법 ($N(0,100^2)$의 사전분포)	18.30	(13.43 , 22.73)

상기 결과를 보면 일양분포와 $N(0,100^2)$의 사전분포하에서 나온 신뢰구간은 거의 차이가 없이 나왔고 Fieller의 방법과 $N(0,10^2)$의 사전분포하에서 구한 신뢰구간이 다른 두개에 비해 약간 짧지만 원래 Fieller의 방법은 보수적이지 못하기 때문에, 즉 실제 신뢰수준이 95%보다 작기 때문에 당연한 결과라 할 수 있으며 다른 하나는 매우 정보적인 사전분포의 영향을 받아 짧아진 것으로 생각된다.

5.2 중도절단자료인 경우

4절의 알고리즘인 Gibbs Sampling 방법을 설명하기 위해 우리가 도입한 중도절단된 자료는 Schmee와 Hahn(1979)에 있는 motorette data에서 따온 것으로 아래 표 5.3에 있다.

표 5.3

150°C	170°C	220°C	x_0°C
	1764	408	408
	2772	408	408
	3444	504	1344
	3542	504	1344
	3780	504	1440
	4860		
	5196		

위 표에서 150°C에서 실험한 10개의 motorettes가 8064시간후에도 모두 고장이 나지 않았고 170°C에서 실험한 10개의 motorettes중 3개는 5448시간후에도 고장이 나지 않았으며 220°C에서 실험한 10개중 5개는 528시간후에도 고장나지 않은 것으로 되어 있고 또 미지의 x_0에서 실험한 10개중 5개는 1680시간 이후에도 고장이 나지 않았다고 하자.

Motorette의 수명을 T, $Y = \log_{10} T$, 온도를 x, $d = 1000(273.2+x)^{-1}$라 할 때 Schmee과 Hahn(1979)에서 처럼 Y를 평균이 $\beta_0+\beta_1 d$인 정규분포의 확률변수로 가정하였다. 4절의 Gibbs Sampling을 이용하여 앞에서 처럼 세 종류의 사전분포인 일양분포, $N(0,10^2)$, $N(0,1000^2)$하에서 i=30, m=50으로 해 x_0에 대한 사후분포를 추정한 결과가 그림 5.2이다. 이 때에도 5.1에서와 같이 i와 m을 정하는 특별한 규칙을 발견하지 못하였고, 여러 값의 i와 m을 실험적으로 시도하여 결과가 수렴된다고 보여지는 값으로 30과 50을 택하게 되었다. 세 종류의 사후분포는 거의 차이가 없어 겹쳐져 있는데 이는 위 자료에서 $k[\beta_j^{(i)}]^2/\sigma_j^{2(i)}$의 값이 사전분포의 분산의 역수인 τ^{-2}에 비해 매우 크게 나오기 때문에 당연한 결과로 판단된다. x_0를 점추정 또는 구간추정하는 기존의 방법을 찾지 못했기 때문에 본 논문의 방법으로만 x_0를 앞의 예에서 처럼 추정한 결과가 표 5.4이다.

표 5.4

사전분포	점추정	95% 신뢰구간
비정보적 일양분포	198.0	(184.5 , 213.1)
$N(0,10^2)$	198.0	(184.5 , 213.0)
$N(0,1000^2)$	198.0	(184.5 , 213.0)

상기 결과를 보면 예상대로 사전분포에 따른 점추정과 신뢰구간은 거의 차이가 없는 것으로 나타났다.

5.3 토의

3절과 4절에서 얻은 사후분포를 두개의 실제 예에 적용한 결과는 비록 정량적으로 보이지는 못하였으나, 현실적으로는 의미있는 수치를 보인 것으로 나타났다. 따라서 3절과 4절에서 제시한 Gibbs Sampling방법을 중도 절단된 자료까지를 포함하는 Calibration문제에 이용하는 것은, 유익할 것이라는 사실을 시사하였다고 결론내릴 수 있다. 한편 차후 연구는 이제 이러한 가능성을 바탕으로 제시된 방법들에 대한 보다 정량적인 평가가 주된 내용이 되리라고 생각한다.

첫째는 제안된 방법이 절단된 자료의 전체자료에 대한 비율에 의하여 어떠한 영향을 받는 지를 Simulation Study를 통하여 고찰하여 보는 것이고, 둘째는 Gibbs Sampling의 구현에 있어서 실제적인 문제인 iteration의 회수 i와 반복회수 m의 결정을 위한 보다 정량화된 방안의 제시이다. 이 연구 또한 자료의 형태에 의존하는 내용이므로 Simulation Study가 요구된다고 본다.

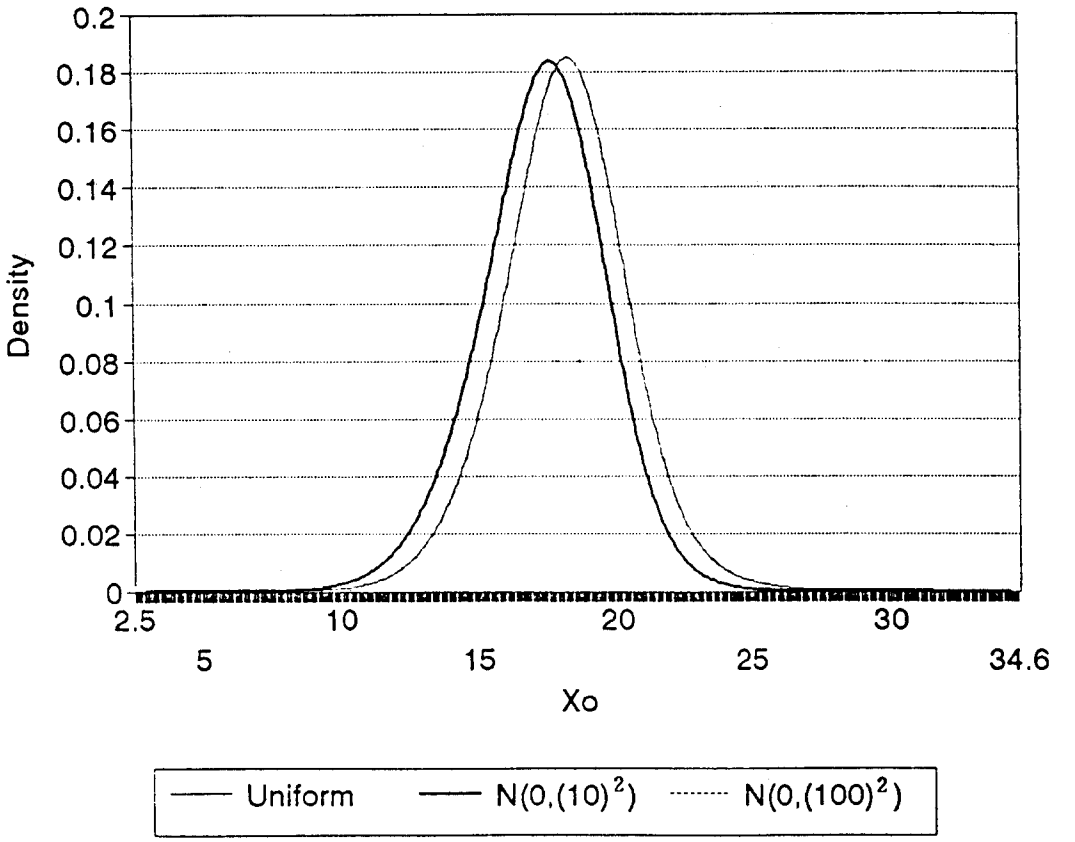

그림 5.1 Posterior density of X_0 for laboratory measurement data

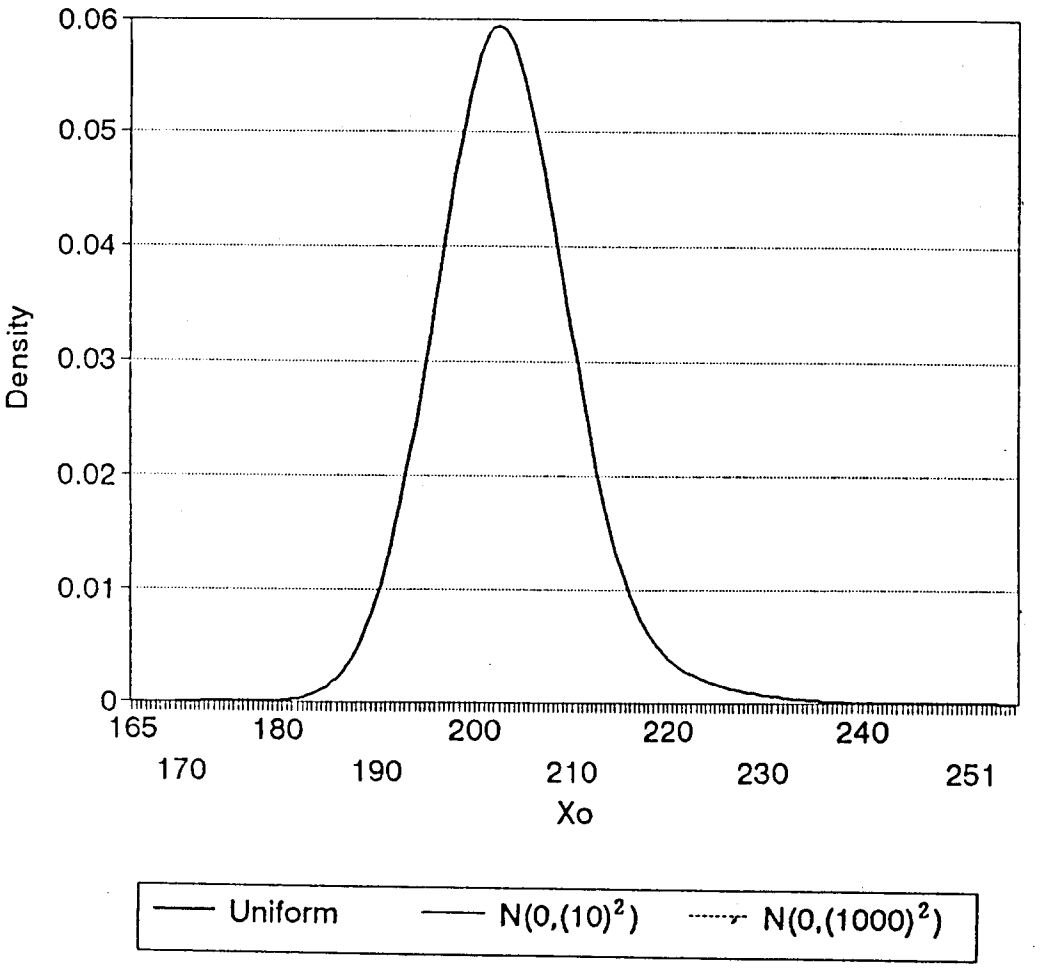

그림 5.2 Posterior density of X_0 for motorette data

참 고 문 헌

[1] Brown, P.J.(1982). Multivariate calibration (with discussion), *Journal of the Royal Statistical society B,* Vol. 44, 287-321.

[2] Buckley, J., and James,I.(1979), Linear regression with censored data, *Biometrika,* Vol. 66, 429-436.

[3] Fieller, E.C.(1932), The distribution of the index in a normal bivariate population, *Biometrika,* Vol. 24, 428-440.

[4] Gelfand, A.E., Hills, S.I., Racine-Poon, A., and Smith,A.F.M.(1990), Illustration of Bayesian inference in normal data models using Gibbs Sampling, *Journal of the American Statistical Association,* Vol. 90, 972-985.

[5] Gelfand, A.E., and Smith, A.f.M.(1990), Sampling based approaches to calculating marginal densities, *Journal of the American Statistical Association,* Vol. 85, 398-409.

[6] Geman,s., and Geman,D. (1984), Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Transations on Pattern Analysis and Machine Intelligence,* Vol. 6, 721-741.

[7] Hill, B.M.(1981), Comment on paper by Hunter and Lamboy, *Technometrics,* Vol. 23, 335-338.

[8] Hoadley, B.(1970), A Bayesian look at inverse regression, *Journal of the American Statistical Association,* Vol. 65, 356-369.

[9] Hunter, W.G. and Lamoy,W.F.(1981), A Bayesian analysis of the linear calibration problem (with discussion), *Technometrics,* Vol. 23, 323-350.

[10] Lwin, T. and Maritz,J.S.(1980), A note on the problem of statistical calibration, *Journal of the Royal Statistical Society,* C, Vol. 29 ,135-141.

[11] Miller, R.G.(1976), Least squares regression with censored data, *Biometrika,* Vol. 63, 449-464.

[12] Naylor, J.C. and Smith,A.F.M.(1982), Application of a method for the efficient computation of posterior distributions, *Applied Statistics,* Vol. 31, 214-225.

[13] Park, N.H. and Lee,S.H.(1990), A Bayesian analysis in multivariate bioassay and multivariate calibration, *Journal of the Korean Statistical Society,* Vol. 19, 71-79.

[14] Rubin, D.(1987), Comment on paper by Tanner and Wong, *Journal of the American Statistical Association,* Vol. 82, 543-546.

[15] Schmee, J. and Hahn,G.J.(1979), A simple method for regression analysis with censored data, *Technometrics,* Vol. 21, 417-432.

[16] Tanner, M.A. and Wong,W.H.(1987), The calculation of posterior distributions by data augmentation (with discussion), *Journal of the American Statistical Association,* Vol. 82, 528-550.

[17] Tierney,L. and Kadane,J.B.(1986), Accurate approximations for posterior moments and marginal densities, *Journal of the American Statistical Association,* Vol. 81, 82-86.

[18] Wei, G.C.G. and Tanner,M.A.(1990),Posterior computations for censored regression data,

Journal of the American Statistical Association, Vol. 85, 829-840.

[19] Zeger, S.L. and Karim,M.R.(1991), Generalized linear models with random effets ; A Gibbs sampling appoach, *Journal of the American Statistical Association,* Vol. 86, 79-86.

On the Calibration Problem with Censored Data[7)]

Nea-Hyun Park[8)], Sukhoon Lee[9)], Nak-Young Lee[10)],
Young-Ok Park[11)], Sang-Ho Lee[12)]

Abstract

This article basically considers the calibration problem with censored data from the Bayesian point of view. The Gibbs sampling method is discussed to solve the difficulty encountered in computing the posterior distribution. Also presented is an approach for impementing the Gibbs sampling in actual data situation with the estimation procedures.

7) This research is supported by the Korean Research Foundation, 1991.
8) Department of Statistics, Chung Nam University, 220 GungDong YousungGu Taejeon, 305-764.
9) Department of Statistics, Chung Nam University, 220 GungDong YousungGu Taejeon, 305-764.
10) Department of Statistics, Chung Nam University, 220 GungDong YousungGu Taejeon, 305-764.
11) Department of Statistics, Chung Nam University, 220 GungDong YousungGu Taejeon, 305-764.
12) Department of Statistics, Chung Nam University, 220 GungDong YousungGu Taejeon, 305-764.