

깁스 표본 기법을 이용한 베이지안 계층적 모형: 야생쥐의 예

송재기 · 이근희 · 하일도

요약. 본 논문은 깁스 표본 기법을 이용하여 Demster et al.(1981)에 의해 소개된 Field Mice 자료를 분석하기 위하여 베이지안 계층적 모형을 적용시켜 보았다. Jeffrey의 사전확률을 이용한 사후 평균을 깁스 표본 기법을 이용하여 구하였고, 이로부터 얻은 베이지안 추정량을 최소 자승 추정량, EM알고리즘을 이용한 랜덤 효과를 포함한 가능도함수에 대한 최대 가능도 추정량(MLR)과 비교하였다.

주제어: EM 알고리즘, 깁스표본기법, 랜덤효과.

1. 서론

계층적 모형(Hierarchical model)은 반복 측정 자료(Repeated measure data 또는 Longitudinal data)로 이루어진 군집된 반응변수의 성격을 묘사해 준다. 이 모형은 의학 및 보건학 분야와 생물학을 비롯한 기초과학 및 공학분야등의 연구에 널리 이용된다. 개체 또는 표본계획의 표본단위에 대한 변이성 때문에 회귀모형에서의 회귀계수는 확률변수인 랜덤효과(Random effect)로 고려될 수 있다. 이러한 회귀계수는 관측이 불가능한 확률변수이지만 계층적 구조를 지닌 통계적모형을 통하여 분석이 가능하다.

계층적모형을 사용한 가능도함수는 형태가 복잡하여 기존의 Newton-Raphson방법으로는 어렵지만, Dempster et al.(1977)이 소개한 EM알고리즘을 이용하여 최대 가능도 추정량을 얻을 수 있다. Jeffrey의 사전 분포를 이용한 베이지안 계층적모형에서는 다차원 적분 계산을 요구하기 때문에 주변 사후 분포를 구할 경우 수치 해석적인 어려움을 겪는다. 이러한 어려움을 극복하기 위하여 본 논문은 통계적 수치 해석적 방법으로 최근에 각광받고 있는 MCMC(Monte-Carlo Markov Chain)방법으로서, 깁스 표본 기법(Gibbs sampler method)을 이용하여 Dempster et al.(1981)이 소개한 Field Mice 자료를 분석하고 EM알고리즘을 이용한 랜덤 효과를 포함한 최대 가능도 추정량(Maximum Likelihood estimator with Random effect; MLR)과 비교하고자 한다.

본 논문의 구성 단계는 다음과 같다. 2절에서는 Field Mice 자료를 소개하고 Dempster et al.

송재기는 경북대학교 통계학과(702-701 대구광역시 북구 산격동 1370)교수이다. 이근희는 서울대학교 통계연구소(151-742 서울특별시 관악구 신림동 산 56-1)의 특별연구원이다. 하일도는 경산대학교 통계학과(712-240 경북 경산시 점촌동 산 75)의 교수이다.

(1981)이 소개한 EM알고리즘을 이용한 Field Mice자료의 분석에 대한 통계적 모형을 언급하였다. 3절에서는 깃스 표본 기법을 언급하였고, 4절에서는 Field Mice자료에 적용되는 베이저안 계층적 모형을 정의하고 분석하였다. 5절에서는 최소 자승 추정량, EM알고리즘을 이용한 랜덤 효과를 포함한 최대 가능도 추정량(MLR)을 깃스 표본 기법을 이용한 사후평균과 비교하였다.

2. 야생쥐 자료 및 MLR

납 제련소(Lead smelter)로부터 떨어진 거리간 인체에 미치는 영향을 조사하기 위하여 다음과 같은 실험을 하였다(Lower and Tsutakawa, 1978). 납 제련소로부터 떨어진 3개 지역에서 야생 쥐를 잡아 꼬리표를 단 후, 비뇨기의 아미노산(Aminolevulinic acid; ALAU)농도를 측정하고 다시 풀어주었다. 그 후, 쥐뿔을 이용하여 야생 쥐를 잡은 후 ALAU농도를 측정하고 다시 풀어주는 과정을 반복하였다. 평균적으로 야생 쥐들은 6일에 한번정도로 잡혔으며, 소수의 야생 쥐들은 한달동안 잡히지 않았기 때문에 그 동안의 ALAU농도의 측정이 불가능하였다. 전문가들에 의하면 시간이 지나면서 ALAU농도가 납 제련소로 인하여 조금씩 감소한다는 사실이 알려져 있다고 한다. 3개 지역은 (그룹1) 납 제련소로부터 0.3km 떨어진 6마리의 야생쥐, (그룹2) 1.7km 떨어진 6마리의 야생쥐, (그룹3) 14.7km 떨어진 8마리 야생 쥐로 구분된다.

이러한 반복 측정 자료(Repeated measures data 또는 Longitudinal data)를 Demster et al. (1981)은 야생 쥐에 대한 ALAU의 로그값과 시간과의 선형관계를 가정한 다음과 같은 통계적 모형을 고려하였다.

$$y_{ijk} = \mu_{1i} + \mu_{2i}x_{ijk} + \gamma_{1ij} + \gamma_{2ij}x_{ijk} + \varepsilon_{ijk}$$

$$i = 1,2,3; \quad j = 1,2, \dots, n_i; \quad k = 1,2, \dots, n_{ij}$$

여기서 y_{ijk} : i 번째 거리 그룹, j 번째 야생쥐, k 번째 시점의 $\log(\text{ALAU}$ 의 농도),

μ_{1i}, μ_{2i} : i 번째 거리 그룹의 모수 효과(fixed effects),

$\gamma_{1ij}, \gamma_{2ij}$: i 번째 그룹 내에서의 j 번째 야생쥐의 랜덤 효과 $\sim BVN(0, \Sigma)$,

x_{ijk} : i, j, k 번째 시간(공변량),

ε_{ijk} : i, j, k 번째 오차 $\sim iid N(0, \sigma^2)$

$(\gamma_{1ij}, \gamma_{2ij})^T, \varepsilon_{ijk}$: 독립.

위와 같은 모형에서의 최대 가능도 추정량은 EM알고리즘을 통하여 구할 수 있으며, Demster et al.(1981)은 이러한 추정량을 MLR(Maximum Likelihood estimator with Random effect)이라고 하였다.

3. 깃스 표본 기법

깃스 표본 기법은 Geman et al.(1984)에 의해 소개된 이미지 재생기법으로 컴퓨터 과학

분야에 널리 쓰여 왔고, 통계학에서는 Gelfand et al.(1990)이 처음 소개하였다. 이 방법은 랜덤효과가 있는 선형모형 분석, 판별분석, 베이지안 분석을 위하여 요구하는 복잡한 함수 적분에 넓은 적용범위를 갖고 있으며 특히 베이지안 추론에 유용하게 사용된다. $\theta_1, \dots, \theta_k$ 를 k 개의 확률변수라 하고 y 를 주어진 자료라 하자. FCD(Full Conditional Density), 즉 k 개의 조건부 사후분포, $p(\theta_i | y, \theta_j, j \neq i), (i = 1, \dots, k)$ 가 알려져 있다고 가정할 경우, 적절한 조건 하에서(Chen, 1993), FCD들을 이용하여 결합분포가 유일하게 결정된다는 사실이 알려져 있다. 깁스 표본 기법은 알려진 FCD들로부터 θ_i 의 주변 사후분포를 구하는 Monte-Carlo방법으로 다음과 같은 절차를 통하여 구현된다.

- 절차1 : 초기치 $\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_k^{(0)}$ 를 정한다.
- 절차2 : $p(\theta_1 | \theta_2^{(0)}, \theta_3^{(0)}, \dots, \theta_k^{(0)}, y)$ 로부터 표본 $\theta_1^{(1)}$ 을 생성
 $p(\theta_2 | \theta_1^{(1)}, \theta_3^{(0)}, \dots, \theta_k^{(0)}, y)$ 로부터 표본 $\theta_2^{(1)}$ 을 생성
 \vdots
 $p(\theta_k | \theta_1^{(1)}, \theta_2^{(1)}, \dots, \theta_{k-1}^{(1)}, y)$ 로부터 표본 $\theta_k^{(1)}$ 을 생성
- 절차3 : t 번 반복 후, 표본 $\theta_1^{(t)}, \theta_2^{(t)}, \dots, \theta_k^{(t)}$ 를 얻는다.

위의 절차를 통하여 형성된 Markov Chain은 다음과 같은 두 가지 유용한 결과를 갖는다.

(결과 1) 분포 수렴: (Geman and Geman, 1984)

$$\lim_{t \rightarrow \infty} p(\theta_1^{(t)}, \dots, \theta_k^{(t)}) = p(\theta_1, \dots, \theta_k | y)$$

(결과 2) Ergodic 정리: (Geman and Geman, 1984; Gelfand and Smith, 1990)

$E(g(\theta_1, \dots, \theta_k) | y)$ 이 존재하는 측도 가능한 함수 g 에 대하여,

$$\frac{1}{t} \sum_{i=1}^t g(\theta_1^{(i)}, \dots, \theta_k^{(i)}) \rightarrow E_{g(y)}(g(\theta_1, \dots, \theta_k) | y) \text{ with probability 1.}$$

위와 같은 절차를 m 번 반복하여 얻은 $(\theta_{1j}^{(i)}, \dots, \theta_{kj}^{(i)}), j = 1, \dots, m$ 을 이용하여, $p(\theta_i | y)$ 의 Rao-Blackwellized 추정값은 다음과 같은 계산식을 통하여 얻을 수 있다.

$$\hat{p}(\theta_i | y) = \frac{1}{m} \sum_{j=1}^m p(\theta_i | y, \theta_{hj}^{(i)}, h \neq i).$$

4. 베이지안 계층적 모형

Field Mice 자료를 분석하기 위하여 다음과 같은 베이지안 계층적 모형을 고려하였다.

$$y_{ijk} = \alpha_{ij} + \beta_{ij} x_{ijk} + \varepsilon_{ijk}$$

$i = 1, 2, 3; \quad j = 1, 2, \dots, n_i; \quad k = 1, 2, \dots, n_{ij}$

여기서 y_{ijk} : i 번째 거리 그룹, j 번째 야생쥐, k 번째 시점의 $\log(ALA U)$ 의 농도,
 α_{ij}, β_{ij} : i 번째 집단 내에서의 j 번째 야생쥐의 랜덤 효과(절편, 기울기),

x_{ijk} : i, j, k 번째 시간(공변량),

ε_{ijk} : i, j, k 번째 오차.

$$\begin{aligned} \text{가정: } \begin{pmatrix} \alpha_{ij} \\ \beta_{ij} \end{pmatrix} &\sim BVN\left(\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix}, \Sigma\right), \Sigma = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{12} & \lambda_{22} \end{pmatrix}; \text{야생쥐들간 조건부 독립,} \\ &(\alpha_i, \beta_i)^T : \text{그룹간 조건부 독립,} \\ &\varepsilon_{ijk} \sim iid N(0, \sigma^2) \text{ 이고 } (\alpha_{ij}, \beta_{ij})^T \text{ 와 독립.} \end{aligned}$$

그리고 모수에 대한 사전 분포들은 독립을 가정한 다음과 같은 Jeffrey의 사전분포를 사용하였다. $p(\alpha_i, \beta_i; \Sigma^{-1}; \sigma^2) = p(\alpha_i, \beta_i)p(\Sigma^{-1})p(\sigma^2)$ 이고 다음과 같은 이변량 정규(BVN), Wishart, 역감마(IG)분포를 갖는다. 즉,

$$\begin{aligned} p(\alpha_i, \beta_i) &= BVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, C_{2 \times 2}\right), \\ p(\Sigma^{-1}) &\propto |\Sigma|^{\frac{1}{2}(q+2)} \text{ (Jeffreys prior), } q = 2 = \dim(\Sigma) \\ p(\sigma^2) &= IG(0, 0) \propto \frac{1}{\sigma^2} \text{ (Jeffreys prior).} \end{aligned}$$

이 경우 가능도 함수는 다음과 같다.

$$\begin{aligned} L(\alpha_i, \beta_i, \Sigma^{-1}, \sigma^2 | y, \alpha_{ij}, \beta_{ij}) \quad (\text{단, } y = \{y_{ijk}\}) \\ &= p(y | \alpha_{ij}, \beta_{ij}, \alpha_i, \beta_i, \Sigma^{-1}, \sigma^2) p(\alpha_{ij}, \beta_{ij} | \alpha_i, \beta_i, \Sigma^{-1}, \sigma^2) \\ &\propto (\sigma^{-2})^{N/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{i,j,k} (y_{ijk} - \alpha_{ij} - \beta_{ij} x_{ijk})^2\right] \\ &\quad \times |\Sigma|^{-M/2} \exp\left[-\frac{1}{2} \sum_{i,j} \begin{pmatrix} \alpha_{ij} - \alpha_i \\ \beta_{ij} - \beta_i \end{pmatrix}^T |\Sigma|^{-1} \begin{pmatrix} \alpha_{ij} - \alpha_i \\ \beta_{ij} - \beta_i \end{pmatrix}\right] \\ &\quad (\text{여기서, } \sum_{i,j,k} 1 = N, \sum_{i,j} = M). \end{aligned}$$

따라서 FCD는 아래와 같이 구해진다.

$$1) p(\sigma^2 | y, \alpha_{ij}, \beta_{ij}, \alpha_i, \beta_i, \Sigma^{-1}) = IG\left(\frac{N}{2}, \frac{1}{2} \sum_{i,j,k} (y_{ijk} - \alpha_{ij} - \beta_{ij} x_{ijk})^2\right) \quad (1)$$

$$2) p\left(\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \middle| y, \alpha_{ij}, \beta_{ij}, \sigma^2, \Sigma^{-1}\right) = BVN\left(\begin{pmatrix} \bar{\alpha}_i \\ \bar{\beta}_i \end{pmatrix}, \frac{1}{n_i} \Sigma\right) \quad (2)$$

단, $\bar{\alpha}_i = \frac{1}{n_i} \sum_j \alpha_{ij}$, $\bar{\beta}_i = \frac{1}{n_i} \sum_j \beta_{ij}$

(여기서 (2)의 FCD를 $C^{-1} = 0$ 으로 놓고 구하였다.)

$$3) p\left(\begin{pmatrix} \alpha_{ij} \\ \beta_{ij} \end{pmatrix} \middle| y, \alpha_i, \beta_i, \sigma^2, \Sigma^{-1}\right) = BVN\left(D_{ij}\left(\frac{1}{\sigma^2} X_{ij}^T Y_{ij} + \Sigma^{-1} \theta_i\right), D_{ij}\right)$$

$$\text{단, } D_{ij} = \left[\frac{1}{\sigma^2} X_{ij}^T X_{ij} + \Sigma^{-1}\right]^{-1},$$

$$X_{ij} = \begin{pmatrix} 1 & x_{ij1} \\ 1 & x_{ij2} \\ \vdots & \vdots \\ 1 & x_{ijn_{ij}} \end{pmatrix}, \theta_i = \begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix}.$$
(3)

$$4) p\left(\Sigma^{-1} \middle| y, \alpha_{ij}, \beta_{ij}, \alpha_i, \beta_i, \sigma^2\right)$$

$$= \text{Wishart}\left[\left[\sum_{i,j} \begin{pmatrix} \alpha_{ij} - \alpha_i \\ \beta_{ij} - \beta_i \end{pmatrix} \begin{pmatrix} \alpha_{ij} - \alpha_i \\ \beta_{ij} - \beta_i \end{pmatrix}^T\right]^{-1}, M-1\right] \text{ 단, } M=20$$
(4)

5. 분석

4절에서 구한 FCD를 이용하여 관심있는 모수 β_i 들의 합에 대한 평균을 구하면서 수렴성을 조사하였다.

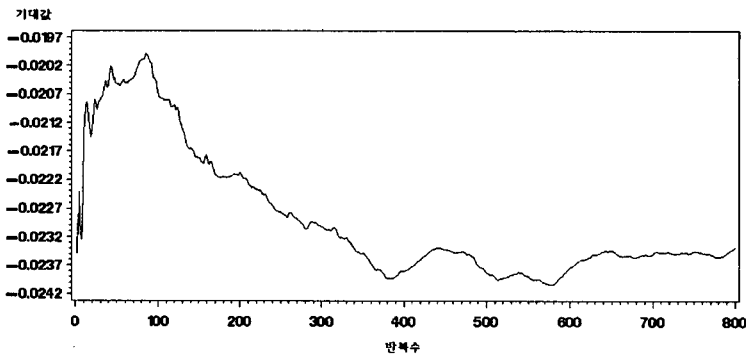


그림 1. 깁스 표본의 수렴성

그림 1에서 보는 바와 같이 600번 반복 후, 다소 안정된 추정량을 얻을 수 있다. 초기값으로는 최소자승 추정량을 사용하였으며, 실제로 각 Markov Chain의 연결회수(반복수)를 1000번으로 정하여 초기값에 의존하지 않도록 하였다. 또한, 독립된 표본을 얻기 위하여, 같은 초기값이지만, 컴퓨터에서 사용하는 Random Number Seed값을 조정하여

3000개의 독립된 Markov Chain을 통하여 3000개의 독립된 표본을 얻었다 (즉, $t=1000$ 이고 $m=3000$). 3000개의 표본을 사용하여 관심있는 모수의 주변 사후 평균을 구하였다. 하나의 예로 σ^2 에 대한 사후 평균 및 사후 분포는 다음과 같이 각각 구할 수 있다.

$$\hat{E}(\sigma^2|y) = \frac{1}{m} \sum_{i=1}^m \sigma_i^{2(i)},$$

$$\hat{p}(\sigma^2|y) = \frac{1}{m} \sum_{j=1}^m p\left(\sigma^2|y, \alpha_{ij}^{(i)}, \beta_{ij}^{(i)}, \alpha_i^{(i)}, \beta_i^{(i)}, \Sigma^{-1(i)}\right)$$

여기서 $t=1000$, $m=3000$ 이고 $\sigma_i^{2(i)}$ 는 t 번 반복 후 (1)에서 생성된 표본이며, $\alpha_{ij}^{(i)}, \beta_{ij}^{(i)}, \alpha_i^{(i)}, \beta_i^{(i)}, \Sigma^{-1(i)}$ 는 t 번 반복 후 (2)~(4)에서 각각 생성된 표본이다.

공분산 성분 에 대한 LSE와 MLR 그리고 깁스 표본 기법에 의한 사후 평균은 다음 표 1과 같다.

표 1에서 GIBBS를 이용한 베이지안 공분산 추정값은 MLR보다 값이 크다는 사실을 발견할 수 있다. 또한, 표 2에서는 그룹 3에 대한 기울기는 거의 0에 가깝고 오히려 양수의 값을 GIBBS에서 제공한다. 그룹 2에 대한 기울기도 MLR에 비하여 현저히 감소되었음을 알 수 있다. 이 사실은 남 제련소로부터 떨어진 거리(그룹효과)에 GIBBS를 통한 베이지안 분석방법이 민감하게 반응한다는 사실을 추측할 수 있다.

표 1. 공분산 성분의 추정값(LSE, MLR, GIBBS)

모수	LSE	MLR	GIBBS
λ_{11}	적용불가	2.321×10^{-2}	4.320×10^{-2}
λ_{12}	적용불가	-3351×10^{-4}	-9.945×10^{-4}
λ_{22}	적용불가	5.194×10^{-6}	24.088×10^{-6}
σ^2	1.101×10^{-1}	1.101×10^{-1}	1.127×10^{-1}

표 2. 각 그룹별 절편 및 기울기의 추정값(LSE, MLR, GIBBS)과 공분산의 추정값

그룹	추정량	절편 (절편의 분산)	기울기 (기울기의 분산)	절편과 기울기의 공분산
그룹 1(0.3km)	LSE	1.244 (3.85×10^{-3})	-12.16×10^{-3} (7.06×10^{-6})	-1.33×10^{-4}
	MLR	1.255 (6.64×10^{-3})	-12.72×10^{-3} (5.98×10^{-6})	
	GIBBS	1.278 (9.65×10^{-3})	-14.87×10^{-3} (0.65×10^{-6})	

그룹 2(1.7km)	LSE	0.997	-9.15×10^{-3}	
	MLR	(2.68×10^{-3})	(5.59×10^{-6})	-1.02×10^{-4}
		0.996	-9.04×10^{-3}	
	GIBBS	(7.55×10^{-3})	(8.56×10^{-6})	-1.97×10^{-4}
0.939		-4.57×10^{-3}		
그룹 3(14.7km)	LSE	0.661	-2.32×10^{-3}	
	MLR	(2.59×10^{-3})	(4.77×10^{-6})	-0.88×10^{-4}
		0.646	-2.11×10^{-3}	
	GIBBS	(5.38×10^{-3})	(5.06×10^{-6})	-1.24×10^{-4}
0.617		0.39×10^{-3}		
		(8.28×10^{-3})	(0.15×10^{-6})	-0.72×10^{-4}

아울러 그림 2, 그림 3, 그림 4로 부터 LSE는 산포가 심하고 MLR과 GIBBS는 비슷한 형태의 기울기와 절편이 선형관계를 이루며 군집된 형태로 그룹을 통하여 나누어진다. 즉 Field Mice자료 분석 방법으로서, MLR과 GIBBS를 이용한 베이지안 접근법은 거리 그룹간의 효과를 쉽게 탐지할 수 있다.

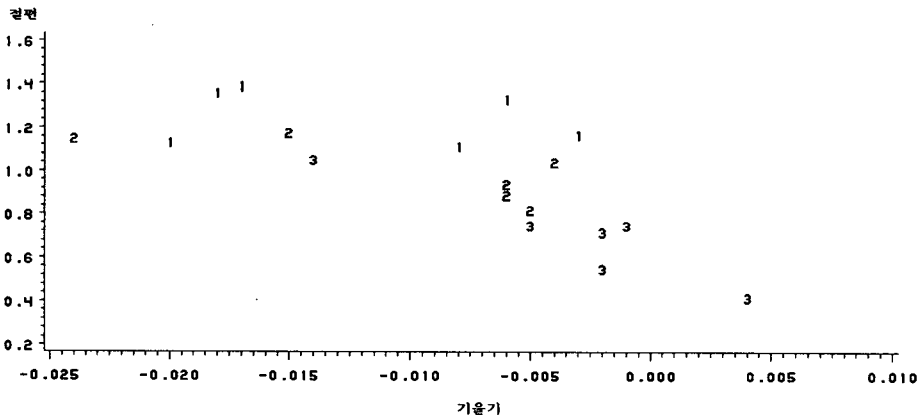


그림 2. 최소제곱추정량의 절편과 기울기(그룹 1: 1, 그룹2: 2, 그룹3:3)

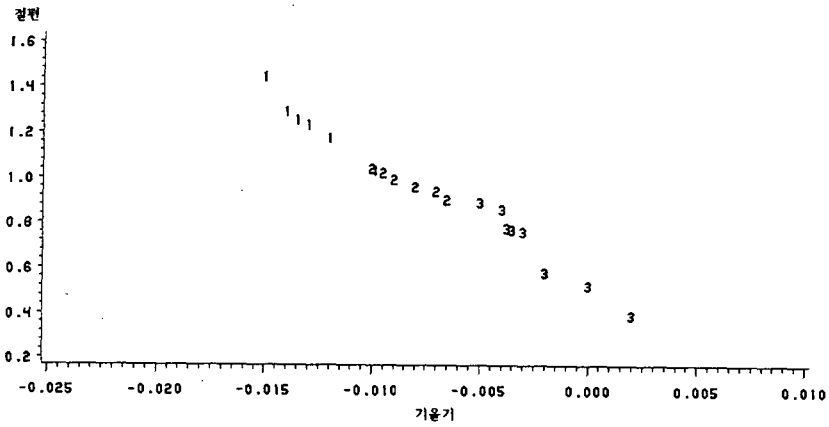


그림 3. MLR의 절편과 기울기(그룹 1: 1, 그룹2: 2, 그룹3:3)

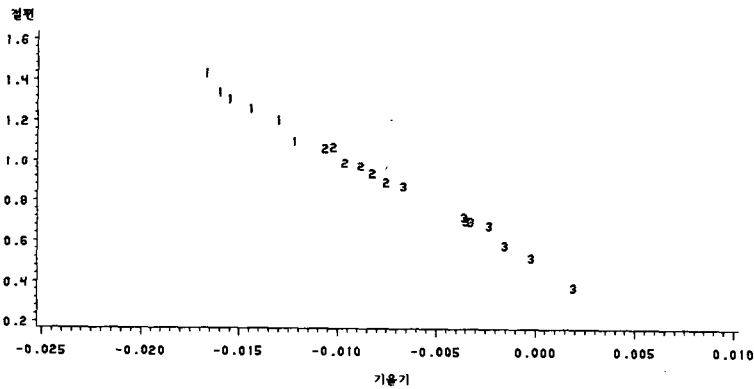


그림 4. Gibbs에 의한 기울기와 사후평균(그룹 1: 1, 그룹2: 2, 그룹3:3)

6. 결론

20마리의 야생 쥐에 대한 분석은 개개의 야생 쥐에 대한 특성으로 인한 변이성오차항에 풀링함으로써 그 값이 커지게 되어(Over-dispersion problem) 최소 자승 추정법의 적용이 불가능하다. 이러한 문제를 극복하기 위하여 베이지안 계층적모형을 사용하였으며, Depm-

ster et al.(1981)이 사용한 MLR과 비교하였다. 베이지안 계층적 모형은 반복 측정 자료에 적용될 수 있는 통계적 모형으로 개개의 야생 쥐에 대한 특성으로 인한 변이성을 모형에 포함시킴으로써 오차항을 감소시키고 그룹사이의 효과를 극대화할 수 있는 분석방법이다. Field Mice자료에서의 베이지안 계층적모형은 그룹간의 차이에 매우 민감하였으며, MLR과 비슷한 추정값을 제공하였다.

참고문헌

- Chen, K. S., (1993) *Asymptotic behavior of Gibbs Sampler*, Journal of the American Statistical Association, Vol. 88, 302-326.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977) *Maximum Likelihood From Incomplete Data via the EM algorithm(with discussion)*, Journal of the Royal Statistical Society, Ser. B, 39, 1-38.
- Dempster, A. P., Rubin, D. B., and Tsutakawa, R. K. (1981) *Estimation in Covariance Components models*, Journal of the American Statistical Association, Vol 76, 341-353.
- 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.
- Geman, S., and Geman ,D. (1984) *Stochastic Relaxation Gibbs Distribution and Bayesian Restoration of Images*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 6, 721-741.
- Lower, W. E., and Tsutakawa, R. K. (1978) *Statistical Analysis of Uniary Aminolevulinic Acid Excretion in the White Footed Mouse Associated With Lead Smelting*, Journal of Environmental Pathology and Toxicology, 1, 551-560.
- Rubin, D. (1987) "Comment on paper by Tanner and Wang", Journal of the American Statistical Association, Vol. 82, 543-546.
- Tanner, M. and Wang, W, H. (1987) *The Calcuration of Posterior Distributions by Data Augmentation (with dicussion)*, Journal of the Royal Statistical Society, Vol. 81, 82-86.

Bayesian Hierarchical Model using Gibbs Sampler Method: Field Mice Example

Jae-Kee Song¹·Gun-Hee Lee²·Il-Do Ha³

Abstract In this paper, we applied bayesian hierarchical model to analyze the field mice example introduced by Demster et al.(1981). For this example, we use Gibbs sampler method to provide the posterior mean and compared it with LSE(Least Square Estimator) and MLR(Maximum Likelihood estimator with Random effect) via the EM algorithm.

Jae-Kee Song is a pofessor of the Dept. of Statistics, Kyungpook National University, Taegu, 702-701, Korea. Gun-hee is a special researcher of Statistical Research Institute, Seoul National University, Seoul, 151-742, Korea. Il-Do Ha is a professor of Dept. of Computer and Statistics, Kyungsan University, Kyungpook , 712-240, Korea.