The Korean Communications
in Statistics Vol. 5, No. 2, 1998
pp. 431-437

경시적 자료의 계층적 베이즈 분석[1)]

김 달 호[2)], 신 임 희[3)]

요 약

본 논문의 목적은 계층적 베이즈 일반화 선형모형을 이용하여 경시적 자료를 분석하는 것이다. 구체적으로 계층적 베이즈 변량효과 모형을 소개하고 무정보적 사전분포 하에서 사후분포가 진(proper)인지에 대한 충분조건을 찾는다. 또한, 깁스(Gibbs) 표본자를 사용하여 제안된 계층적 베이즈 절차의 수행에 관해 논의한다. 현실자료를 사용하여 제안된 계층적 베이즈 분석을 예시하고, 이에 대응하는 경험적 베이즈 분석과 비교한다.

1. 서 론

통계학에서 가장 기본적인 문제중 하나는 주어진 자료를 바탕으로 그것을 가장 잘 설명할 수 있는 새로운 모형을 개발하고 그 모형을 적용함으로서 새로운 분석방법을 제공하는 것이다. 본 논문에서는 최근에 매우 활발하게 연구되고 있는 경시적 자료에 대한 통계적 모형으로 계층적 베이지안 모형을 제안하고 깁스(Gibbs) 표본자를 사용한 통계적 추론을 연구하고자 한다. 경시적 자료의 frequentist 분석에 대한 통계적 모형과 방법론적 연구 결과는 최근 Diggle, Liang 및 Zeger (1994)에 의해 매우 잘 요약되었다.

경시적 자료는 각 개체를 시간에 따라 반복하여 측정한 관측치들로 구성된다. 각 개체가 단 한 번 측정되는 횡단면연구(cross-sectional study)와는 달리, 경시적 연구(longitudinal study)는 개체들이 시간에 따라 반복 측정됨으로 개체내의 시간에 따른 변화(aging effect)를 찾아낼 수 있다. 경시적 자료에는 개체내의 반복된 측정치들 사이에 자연스럽게 상관(natural correlation)이 생기게 되고, 경시적 자료에 대한 어떤 통계적 분석도 이 상관을 반드시 고려해야 한다.

최근 일반화 선형모형(generalized linear model; GLM)에 대한 연구의 활성화와 더불어 이산형 반응변수를 가진 경시적 자료에 대한 분석으로 일반화 선형모형을 확장한 모형들이 많이 연구되었다. 구체적으로 주변모형(marginal model), 변량효과 모형(random effects model) 그리고 추이모형(transition model)등 이다.

위의 모형들에 관한 frequentist 추론으로 조건부 최우추정법(conditional maximum likelihood estimation), 최우추정법(maximum likelihood estimation), 일반화 추정식(generalized estimating

1) 이 논문은 1997년 한국학술진흥재단의 공모과제 연구비에 의하여 연구되었음.
2) (702-701) 대구광역시 북구 산격동 1370번지 경북대학교 통계학과 조교수.
3) (705-034) 대구광역시 남구 대명4동 3056-6번지 대구효성가톨릭대학교 의학과 전임강사.

equation; GEE)등이 많이 사용되었다. (Diggle, Liang 및 Zeger (1994) 참조). 한편, Waclawiw와 Liang(1994)는 변량효과 모형에 대하여 경험적 베이즈(empirical Bayes; EB) 추론을 연구하였다.

본 논문의 목적은 경시적 자료의 분석을 위하여 계층적 베이즈(hierarchical Bayes; HB) 일반화 선형모형(GLM)을 도입하여 frequentist 추론과 경험적 베이즈 추론에 대응하는 완전히 베이지안적(fully Bayesian) 추론을 제안하는 것이다. 여기서 우리의 관심은 변량효과 모형에 대한 HB 분석이다.

본 논문의 대략적 개괄은 다음과 같다. 2절에서 일반적인 계층적 베이즈 변량효과 모형을 도입하고, 비정보적(noninformative) 사전분포 하에서 사후분포가 진(proper)일 충분조건을 찾는다. 3절에서 현실 자료를 사용하여 계층적 베이즈 분석을 예시하고, Waclawiw와 Liang(1994)의 경험적 베이즈 분석과 비교한다.

계층적 베이즈 변량모형은 시간에 따른 개체내의 상관을 각 개체에 변량효과를 사용하여 모형화 할 뿐만아니라, 사전분포의 초모수(hyperparameter)에 적절한 분포를 부여함으로서 다른 개체들로 부터 "borrow strength"를 가능하게 한다. 경시적 자료에 계층적 베이즈 분석을 시도하는 기본적 개념은 모형을 계층구조(hierarchy)로 표현하여 비슷한 특성을 가지는 다른 개체들로부터 뿐만 아니라 관심있는 개체에 대해 시간(across time)으로 부터도 "borrow strength"를 할 수 있다는 것이다.

통상 쓰이는 경험적 베이즈 방법에 대한 계층적 베이즈 방법의 장점은 계층적 베이즈 방법이 경험적 베이즈 방법보다 더욱 신뢰할 만한 표준오차를 제공한다는 것이다. 실제로 경험적 베이즈 방법으로 추정된 사후분산은 초모수의 추정에 기인하는 추가변이를 제대로 반영하지 못한다. 왜냐하면 계층적 베이즈 방법은 초모수의 불확실성(uncertainty)을 분포를 통해 모형화하는 반면 경험적 베이즈 방법은 초모수에 단순히 점추정치를 대입하기 때문이다.

2. 계층적 베이즈 변량효과 모형

Y_{ij}를 j번째 시간에서 i번째 개체의 반응이라 하자. ($j=1,\ldots,n_i; i=1,\ldots,m$). 주어진 θ_{ij}에 대하여, Y_{ij}가 독립이면서 다음과 같은 밀도함수를 가지는 일반화 선형모형(GLM)을 생각하자.

$$f(y_{ij}|\theta_{ij}) = \exp[\{y_{ij}\theta_{ij} - \psi(\theta_{ij})\}/\phi_{ij}]h(y_{ij};\phi_{ij}). \tag{2.1}$$

여기서 $\phi_{ij}(>0)$는 안다고 가정한다. 예를 들면 이항분포, 포아송분포, 분산이 알려진 정규분포등이 이에 속한다. (2.1)에서 주어진 모형에 대하여

$$E(Y_{ij}|\theta_{ij}) = \psi'(\theta_{ij});\ V(Y_{ij}|\theta_{ij}) = \phi_{ij}\psi''(\theta_{ij}) \tag{2.2}$$

이 성립한다. 여기서 $\psi'(\cdot)$은 증가함수이다. 또한 θ_{ij}는 흔히 GLM의 정준모수(canonical parameter)로 일컬어 진다.

흔히 θ_{ij}는 다음과 같이 모형화 된다.

$$h(\theta_{ij}) = \boldsymbol{x}_{ij}^T \boldsymbol{b} + u_i \tag{2.3}$$

여기서 h는 엄격한 증가함수이고, $\boldsymbol{b}(p\times 1)$는 알려지지 않은 회귀계수의 벡타이고, u_i는 i번째 개체와 관련된 변량효과(random effect)이다. 또한 $u_i \sim iid\ N(0,\sigma_u^2)$라고 가정한다. 계층적 베이즈 모형을 설정하기 위해서, $\boldsymbol{b}$ 와 σ_u^2가 주변적으로 독립(marginally independent)이고, $\boldsymbol{b} \sim Unif(R^p)$ 그리고 σ_u^2가 역감마분포 IG($\frac{1}{2}a, \frac{1}{2}b$)를 따른다고 가정하자. 여기서 IG($a,b$)는 밀도함수가 $f(x) \propto x^{-(b+1)}e^{-a/x}$ 형태임을 의미한다.

위의 변량모형은 Breslow와 Clayton(1993) 그리고 Zeger와 Karim(1991)에 의해서 관련된 그러나 다른 상황에서 사용되었다. Zerger와 Karim(1991)은 경시적 자료가 아닌 다른 문제에서 비정보적 사전분포(noninformative prior)를 고려하였다. 그러나 Zerger와 Karim(1991)에 의해 사용된 사전분포가 진 사후분포(proper posterior)에 과연 이르게 하는지를 확인할 수 없는 실정이다. 이항분포의 경우 Natarajan과 McCulloch(1995)는 사후분포가 진일 필요충분조건을 제시하였다. 최근 Ghosh, Natarajan, Stroud 그리고 Carlin(1998)에 의해 소영역 추정(small area estimation)에 사용된 GLM 모형에서 진 사후분포를 가질 충분조건이 주어졌다.

여기서 $\boldsymbol{y}=(y_{11},\ldots,y_{1n_1},\ldots,y_{m1},\ldots,y_{mn_m})^T$, $\boldsymbol{\theta}=(\theta_{11},\ldots,\theta_{1n_1},\ldots,\theta_{m1},\ldots,\theta_{mn_m})^T$ 그리고 $\boldsymbol{X}=(\boldsymbol{x}_{11},\ldots,\boldsymbol{x}_{1n_1},\ldots,\boldsymbol{x}_{m1},\ldots,\boldsymbol{x}_{mn_m})$라 하자. 또한 rank($\boldsymbol{X}$) = p라고 가정하자.

다음 정리는 위에서 서술된 경시적 자료에 대한 계층적 베이즈 일반화 선형 혼합모형(HB GLMM)하에서 주어진 $\boldsymbol{y}$에 대해 $\boldsymbol{\theta}$, $\boldsymbol{b}$ 그리고 σ_u^2의 결합 사후분포(joint posterior)가 진(proper)이 될 충분조건을 제시한다. 여기서 $\theta_{ij}\in(\underline{\theta}_{ij}, \overline{\theta}_{ij})$ $(j=1,\ldots,n_i;\ i=1,\ldots,m)$라 하자.

정리 1. $a>0$이라 하자. $f(y_{ij}\mid\theta_{ij})$가 모든 θ_{ij}에 대해 유계(bounded)이고 그리고 적어도 하나의 짝 (i,j)에 대하여

$$I_{ij}=\int_{\underline{\theta}_{ij}}^{\overline{\theta}_{ij}} \exp[(y_{ij}\theta_{ij}-\psi(\theta_{ij}))/\phi_{ij}]h'(\theta_{ij})d\theta_{ij} < \infty \tag{2.4}$$

이 성립한다고 가정하자. 또한, $\sum_{i\in S} n_i + b > p$라고 하자. 여기서 $S=\{i : I_{ij} < \infty$ 적어도 하나의 j에 대해 $\}$이다. 그러면, 주어진 $\boldsymbol{y}$에 대해 $\boldsymbol{\theta}$, $\boldsymbol{b}$ 그리고 σ_u^2의 결합 사후분포가 진(proper)이다.

<증명> 주어진 $\boldsymbol{y}$에 대해 $\boldsymbol{\theta}$, $\boldsymbol{b}$ 그리고 σ_u^2의 결합 사후분포는 다음과 같이 주어진다.

$$\Pi(\boldsymbol{\theta},\boldsymbol{b},\sigma_u^2\mid\boldsymbol{y}) \propto \prod_i\prod_j \exp[\phi_{ij}^{-1}(y_{ij}\theta_{ij}-\psi(\theta_{ij})]$$

$$\times(\sigma_u^2)^{-\sum_{i=1}^{m} n_i/2}\exp[-\sum_i\sum_j(h(\theta_{ij})-\boldsymbol{b}^T\boldsymbol{x}_{ij})^2/(2\sigma_u^2)]\prod_i\prod_j h'(\theta_{ij})$$
$$\times(\sigma_u^2)^{-b/2-1}\exp(-a/(2\sigma_u^2)).$$

먼저 (2.4)에 주어진 I_{ij}가 무한인 θ_{ij}에 대해서 적분을 한다. $\boldsymbol{\theta}^*$를 I_{ij}가 유한인 θ_{ij}들의 모임이라 하자. 가정에 의해서 $\boldsymbol{\theta}^*$는 적어도 한 원소를 가진다. 그러면, 주어진 $\boldsymbol{y}$에 대해 $\boldsymbol{\theta}^*$, $\boldsymbol{b}$ 그리고 σ_u^2의 결합 사후분포는 다음과 같이 주어진다.

$$\Pi(\boldsymbol{\theta}^*, \boldsymbol{b}, \sigma_u^2 \mid \boldsymbol{y}) \propto \prod\prod_{(i,j):I_{ij}<\infty}\exp[\phi_{ij}^{-1}(y_{ij}\theta_{ij}-\psi(\theta_{ij})]$$
$$\times(\sigma_u^2)^{-\sum_{i\in S} n_i/2}\exp[-\sum\sum_{(i,j):I_{ij}<\infty}(h(\theta_{ij})-\boldsymbol{b}^T\boldsymbol{x}_{ij})^2/(2\sigma_u^2)]$$
$$\times\prod\prod_{(i,j):I_{ij}<\infty}h'(\theta_{ij})(\sigma_u^2)^{-b/2-1}\exp(-a/(2\sigma_u^2)).$$

먼저 $\boldsymbol{b}$에 대해서 적분한 다음에 σ_u^2에 대해서 적분한다. 그러면 다음과 같이 주어진다.

$$\pi(\boldsymbol{\theta}^*\mid\boldsymbol{y}) \le K\prod\prod_{(i,j):I_{ij}<\infty}\exp[\phi_{ij}^{-1}(y_{ij}\theta_{ij}-\psi(\theta_{ij}))]h'(\theta_{ij}),$$

여기서 $K(>0)$는 θ_{ij}에 의존하지 않는 상수이다.

따라서 가정에 의해서 $\int\pi(\boldsymbol{\theta}^*\mid\boldsymbol{y})\,d\boldsymbol{\theta}^*<\infty$이다. □

정리 1에 의하면 구체적으로 $Y_{ij}\sim Bin(n_{ij}, \exp(\theta_{ij})/(1+\exp(\theta_{ij})))$이면 적어도 한 짝 (i,j)에 대해 $1\le y_{ij}\le n_{ij}-1$이어야 한다. 그리고 만약 $Y_{ij}\sim Poisson(\exp(\theta_{ij}))$이면 적어도 한 짝 (i,j)에 대해 $y_{ij}>0$이어야 한다.

조건 (2.4)를 적절히 변형시킴으로 임의의 함수 g에 대해 (흔히 $g=\psi'$) $g(\theta_{ij})$의 적률이 유한함을 보일 수 있다. 예컨대, (2.4) 대신에 다음의 조건이 만족되면

$$I_{ij}=\int_{\underline{\theta}_{ij}}^{\overline{\theta}_{ij}}|g(\theta_{ij})|^r\exp[(y_{ij}\theta_{ij}-\psi(\theta_{ij}))/\phi_{ij}]h'(\theta_{ij})d\theta_{ij}<\infty$$

$g(\theta_{ij})$의 r번째 적률이 유한하다.

위의 HB GLMM모형에서의 통계적 추론은 다차원 적분문제의 어려움이 있지만 최근에 베이지안 추론에서 널리 쓰이는 Gibbs 표본방법에 의해서 해결이 가능하다. Gibbs 표본방법을 수행하기 위한 조건부 밀도함수를 다음과 같이 구할 수 있다.

(i) $\boldsymbol{b}\mid\boldsymbol{\theta},\sigma_u^2,\boldsymbol{y}\sim N[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{h}(\boldsymbol{\theta}),\ \sigma_u^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}]$

여기서 $\boldsymbol{h}(\boldsymbol{\theta})=(h(\theta_{11}),\ldots,h(\theta_{mn_m}))^T$;

(ii) $\sigma_u^2\mid\boldsymbol{\theta},\boldsymbol{b},\boldsymbol{y}\sim IG\left(\frac{1}{2}\left\{a+\sum_{i=1}^{m}\sum_{j=1}^{n_i}(h(\theta_{ij})-\boldsymbol{x}_{ij}^T\boldsymbol{b})^2\right\},\ \frac{1}{2}\left(\sum_{i=1}^{m}n_i+b\right)\right)$;

(iii) $\pi(\theta_{ij} \mid \theta_{kl}((k,l)\neq(i,j)),\ \boldsymbol{b}, \sigma_u^2,\ \boldsymbol{y})$
$\propto \exp[\{y_{ij}\theta_{ij}-\psi(\theta_{ij})\}/\phi_{ij}]\exp[-(h(\theta_{ij})-\ \boldsymbol{x}_{ij}{}^T\boldsymbol{b})^2/(2\sigma_u^2)].$

여기서 (i) 과 (ii)에 주어진 조건부 밀도함수로 부터는 쉽게 표본을 추출할 수 있지만, (iii)의 경우는 표본을 쉽게 추출할 수 있는 표준 밀도함수가 아니다. 그러나 이러한 어려움은 Metropolis 기각 알고리즘을 사용함으로 극복될수 있다. 한편, 다른 대안으로 $\pi(\theta_{ij} \mid \cdot)$이 모두 대수를 취했을 때 오목하기(log-concave) 때문에 Gilks 와 Wild(1992)의 ARS(adaptive rejection sampling) 알고리즘을 사용할 수도 있다.

3. 예 제

이 절에서는 Jones와 Kenward(1989)에 주어진 현실 자료에 대한 계층적 베이즈 분석을 예시한다. 주어진 자료는 뇌혈관의 결함에 관해 두 처치(treatment)와 두 기간(period)을 가진 교차실험(cross-over trial)의 결과이다. 동일 자료에 관한 경험적 베이즈 분석이 Waclawia와 Liang(1994)에 주어져 있다.

우리의 관심은 진짜약(active drug)과 위약(placebo)의 뇌혈관 결함에 대한 효과를 심전도 결과로 측정하여 비교하는 것이다. 고려 대상인 67명의 환자중에서 임의로 선택된 34명에게 진짜약을 먼저 그리고 위약을 나중에 투여하고 나머지 33명에게는 위약을 먼저 그리고 진짜약을 나중에 투여한다.

Y_{ij}는 i번째 환자의 j번째 관측치로서 심전도 결과가 정상이면 1 비정상이면 0으로 나타낸 이항반응이라고 하자. ($i=1,\ldots,67;\ j=1,2$). 공변량으로서 처치(TMT)는 위약(B)에 대해서는 1, 진짜약(A)에 대해서는 0으로 나타낸다. 그리고 또 하나의 다른 공변량으로 기간(PER)은 기간 1에 대해서는 0, 기간 2에 대해서는 1로 나타낸다. 그러면 다음과 같은 계층적 모형을 생각할 수 있다.

$$y_{ij} \mid p_{ij} \sim Bernoulli(p_{ij});$$

$$\theta_{ij} = \text{logit}\,(p_{ij}) = b_0 + b_1 x_{1ij} + b_{2ij} + u_i,$$

여기서 $x_{1ij}=\begin{cases}1 & \text{위약}(B)\\ 0 & \text{진짜약}(A)\end{cases}$ 그리고 $x_{2ij}=\begin{cases}1 & \text{기간2}\\ 0 & \text{기간1}\end{cases}$ 이다. 처리와 기간의 교호작용은 frequentist 로지스틱 회귀분석과 베이지안 분석 모두에서 유의하지 않은 것으로 판명되어 위의 모형에 포함되지 않았다. 변량효과 u_i들은 독립인 $N(0,\ \sigma_u^2)$ 분포를 따른다.

계층적 모형을 완성하기 위해서 $\boldsymbol{b}=(b_0, b_1, b_2)^T$에 R^3상의 일양 사전분포를 그리고 σ_u^2에 IG($\frac{a}{2}$,0) 사전분포를 부여한다. 여기서 초사전분포(hyperprior)를 near diffuse화 시키기 위해서 $b=0$ 그리고 a를 0에 가까운 아주 작은 값으로 취한다. a의 값에 대한 추론의 민감도(sensitivity)를 조사하기 위해 a의 값으로 .02, .002, .0002를 고려한다. 위와 같이 IG분포의 모수

를 택하는 근거는 정리 1에 의하면 주어진 자료의 경우 $\sum_{i\in S} n_i = 12$이므로 b는 $12+b > 3$을 만족해야 하고 $a > 0$이어야 한다. 한편 a, b를 작은 값으로 택하는 이유는 초사전분포를 무정보적으로 만들기 위해 diffuse화 시킬 필요가 있기 때문이다. 표 2에 의하면 우리의 분석이 a값에 상대적으로 그렇게 민감하지 않다.

표 1에 회귀계수 b_0, b_1, b_2와 분산성분 σ_u^2의 점추정값들이 괄호안에 표준오차와 함께 주어져 있고, equal tail 95% 신뢰구간도 주어져 있다. 여기서 HB 분석은 a가 0.002인 경우의 값들이다. 비교를 위하여 Waclawiw와 Liang(1994)의 EB 분석결과도 함께 소개했다. 기대했던대로 HB 및 EB 점추정값들은 아주 비슷하다. 그러나 naive EB 방법이 표준오차를 흔히 과소추정(under-estimation) 하기 때문에, Waclawiw와 Liang(1994)은 이러한 단점을 보완하기 위해 EB 절차의 표준오차에 대한 붓스트랩 추정치를 사용했으나 이것 역시 HB 방법에 의한 추정치보다 항상 적음을 알 수 있다. 따라서 이 문제에 대해 붓스트랩 방법에 의해서도 EB 절차의 표준오차의 과소추정은 해결되지 않음을 알 수 있다.

<표 1> b_0, b_1, b_2와 σ_u^2의 HB 및 EB 추정치(표준오차) 그리고 Equal Tail 95% 신뢰한계

모수	추정치		95% HB 신뢰한계		95% EB 신뢰한계	
	HB	EB	아래	위	아래	위
b_0	1.483 (0.420)	1.680 (0.369)	0.679	2.336	0.957	2.403
b_1	-0.683 (0.462)	-0.778 (0.435)	-1.590	0.213	-1.631	0.074
b_2	-0.329 (0.455)	-0.417 (0.385)	-1.227	0.580	-1.171	0.337
σ_u^2	1.030 (0.199)	2.370 (0.147)	0.698	1.474	2.082	2.658

<표 2> σ_u^2에 대한 IG($\frac{a}{2}$, 0) 사전분포에 대한 HB 추정치(표준오차)의 민감도 조사

모수	$a=0.02$	$a=0.002$	$a=0.0002$
b_0	1.482 (0.419)	1.483 (0.420)	1.483 (0.420)
b_1	-0.682 (0.462)	-0.683 (0.462)	-0.683 (0.463)
b_2	-0.329 (0.455)	-0.329 (0.455)	-0.330 (0.455)
σ_u^2	1.030 (0.199)	1.030 (0.199)	1.030 (0.199)

참 고 문 헌

[1] Breslow, N.E. and Clayton, D.G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, 88, 9-25.

[2] Diggle, P.J., Liang, K.Y. and Zeger, S.L. (1994). *Analsis of Longitudinal Data.* Oxford University Press, Oxford.

[3] Ghosh, M., Natarajan, K., Stroud, T.W.F. and Carlin, B.P. (1998). Generalized Linear Models for Small Area Estimation. Journal of the American Statistical Association, 93, 273-282.

[4] Gilks, W.R. and Wald, P. (1992). Adaptive Rejection Sampling for Gibbs Sampling. *Applied Statistics*, 41, 337-348.

[5] Jones, B. and Kenward, M.G. (1989). *Design and Analysis of Cross Over Trials*, Chapman and Hall, London.

[6] Natarajan, R. and McCulloch, C.E. (1995). A Note on the Existence of the Posterior Distribution for a Class of Mixed Models for Binomial Responses. *Biometrika*, 82, 639-643.

[7] Waclawiw, M.A. and Liang, K.Y. (1994). Empirical Bayes Estimation and Inference for the Random Effects Model with Binary Response. *Statistcs in Medicine*, 13, 541-551.

[8] Zeger, S.L. and Karim, M.R. (1991). Generalized Linear Models with Random Effects; a Gibbs Sampling Approach. Journal of the American Statistical Association, 86, 79-86.