응용통계연구 제 15권 1호, 2002년, pp. 73-84

제한조건이 있는 선형회귀 모형에서의 베이지안 변수선택*

오만숙[1)]

요 약

계수에 대한 부등 제한조건이 있는 선형 회귀모형은 경제모형에서 가장 흔하게 다루어지는 것 중의 하나이다. 이는 특정 설명변수에 대한 계수의 부호를 음양 중 하나로 제한하거나 계수들에 대하여 순서적 관계를 주기 때문이다. 본 논문에서는 이러한 부등 제한이 있는 선형회귀 모형에서 유의한 설명변수의 선택을 해결하는 베이지안 기법을 고려한다. 베이지안 변수선택은 가능한 모든 모형의 사후확률 계산이 요구되는데 본 논문에서는 이러한 사후확률들을 동시에 계산하는 방법을 제시한다. 구체적으로, 가장 일반적인 모형의 모수에 대한 사후표본을 깁스 표본기법을 적용시켜 얻은 후 이를 이용하여 모든 가능한 모형의 사후확률을 계산하고 실제적인 자료에 본 논문에서 제안된 방법을 적용시켜 본다.

주요용어: 정규선형 모형, 사후확률, 마코브체인 몬테칼로

1. 서론

본 논문에서는 모수에 대한 부등 제한조건이 있는 정규선형모형에서, 변수선택, 즉 모형선택의 문제를 다루고자 한다. 일반적으로 널리 쓰이는 검정법으로 우도비 검정법 (likelihood ratio test)은 최우 추정치에 근거한다. 그러나 모수에 제한 조건이 있는 선형회귀 모형의 경우 최우 추정치가 주어진 제한조건을 만족시키지 못하는 경우가 종종 발생하므로 적용에 문제가 발생한다.

이러한 우도비 검정의 문제점을 해결하는 방안으로 베이지안 접근법을 고려할 수 있다. 베이지안 모형선택은 각 모형의 사후확률 혹은 그들의 비율(ratio)인 베이즈 상수 (Bayes Factor) 에 의존하는데 각 모형의 사후확률은 적분우도함수 (integrated likelihood)에 비례한다. 따라서 모든 가능한 모형에 대하여 적분우도함수를 계산하여 이들을 비교하여야 한다. 모수에 제한조건이 없는 선형회귀 모형은 적분우도함수가 수리적으로 쉽게 구해지므로 베이지안 모형선택에 어려움이 없다. 그러나 모수에 부등 제한조건이 있는 경우에는 적분 우도함수를 구하기 위한 적분이 수리적으로 수행되지 않는 계산상의 어려움이 따른다. 본 논문에서는 이를 마코브 체인 몬테칼로(MCMC) 기법의 일종인 깁스 표본기법으로부터 얻어지는 모수의 사후표본을 사용하여 해결하고자 한다. 깁스 표본기법은 다차원 모수의

* This work was supported by Grant #99-N6-01-01-A-03 for the Women's University from Korea Ministry of Science and Technology.

1) (120-750) 서울시 서대문구 대현동 21 이화 여자 대학교 통계학과, 부교수
E-mail: msoh@mm.ewha.ac.kr.

사후 표본을 생성하는데, 모수의 원소모수 (component parameter)의 조건부 사후분포로부터 반복적으로 (iterative) 표본을 생성하는 방법이다. 깁스 표본기법은 저차원 분포로부터의 표본 생성으로 구성되기 때문에 복잡한 제한조건도 쉽게 처리할 수 있다는 장점이 있어 본 논문에서 고려하고자 하는 제한이 있는 선형회귀 모형에 적용하기에 적절한 기법이다.

설명변수가 p 개 있다고 하면 변수선택에서 가능한 모형은 2^p 개 존재한다. 따라서 2^p 개의 모형에 대한 사후확률의 계산이 요구된다. 그런데 모형의 사후확률과 해당 모형의 모수의 사후밀도함수와는 밀접한 관련이 있다. Raftery(1996)에서 보인 바와 같이, 모형의 사후 확률은 사후밀도함수의 정규화 상수 (normalizing constant)에 해당한다. 따라서 모형의 모수의 사후밀도함수를 계산할 수 있다면 해당하는 모형의 사후확률을 쉽게 구할 수 있고 각 모형의 사후확률 추정의 문제는 각 모형에 존재하는 모수의 사후밀도함수를 추정하는 문제로 귀결된다.

그런데 만약 사후확률 계산을 위하여 모든 가능한 2^p개의 모형에 깁스 표본기법을 적용하여 모수의 사후표본을 생성해야 한다면, 이는 p가 아주 작은 경우를 제외하고는 계산적 부담이 매우 클 것이다. 우리는 이 문제를 Oh(1999)가 제안한 사후밀도함수 추정법을 통하여 해결하고자 한다. Oh(1999)가 제안한 방법을 사용하면, 모형에서 얻어진 사후표본을 이용하여 모수벡터의 임의의 부분집합에 대한 사후밀도함수를 동시에 얻을 수 있으며 또한 사후 표본의 생성 외에 추가적인 비용이 거의 없다는 장점이 있다. 따라서 가능한 모든 모형들을 전체 모형(full model) 에 내포된 (nested) 모형들로 나타낼 수 있다면 내포된 모형들의 모수가 바로 전체 모형의 모수의 부분집합이 될 것이고 이에 Oh(1999)의 방법을 적용한다면 모든 가능한 모형의 모수의 사후밀도함수를 동시에 추정할 수 있을 것이다. 이로써 아주 작은 비용으로 모든 가능한 변수선택 모형을 고려할 수 있다.

본 논문은 다음과 같이 구성되어 있다. 2절에서는 부등 제한이 있는 선형회귀 모형과 모수의 사전분포를 제시한다. 3절에서는 모수의 사후표본 생성을 위한 깁스 표본기법을 제시한다. 4절에서는 깁스 표본을 이용한 쉽고 효율적인 베이지안 변수 선택기법을 제시한다. 5절에서는 미시간 대학생들의 임대료 자료를 제안된 기법에 적용시켜 본다.

2. 부등제한이 있는 선형회귀 모형과 사전분포

선형회귀 모형은 다음과 같이 정의된다. i 번째 대상의 반응치를 y_i 라 하고 이에 대응하는 설명변수를 $x_i = (x_{i1}, ..., x_{ip})$라 하면

$$y_i = x_i^t \beta + \epsilon_i$$

이며 ϵ_i는 독립적으로 $N(0, \sigma^2)$분포를 따르고 $\beta = (\beta_1, ..., \beta_p)^t$는 회귀계수를 나타내는 모수이다. 벡터와 행렬을 사용하여

$$y = (y_1, ..., y_n)^t, \quad X = (x_1, ..., x_n)^t,$$

$\epsilon = (\epsilon_1, .., \epsilon_n)^t$ 라 놓으면 위 모형은

$$y = X\beta + \epsilon$$

로 나타낼 수 있다. 본 논문에서는 위 선형모형에서 계수 β 가 주어진 부등제한조건 B를 만족해야 하는 제한조건을 가진 경우를 고려한다.

주어진 모형에서 모수는 β와 σ^2이다. 베이지안 추론을 위해서는 모수의 사전분포를 가정해야 하는데 β의 사전분포로는 $N(\mu_0, \Sigma_0)I_B(\beta)$, 즉, 제한된 정규분포를 가정한다. 여기에서 I는 지시함수로 β가 제한조건 B를 만족시키면 1 아니면 0 이다. 따라서 β의 부등제한조건을 사전분포에 포함시킨 것인데 이러한 사전분포를 사용하면 β의 사후분포는 반드시 B를 만족시킨다는 것은 널리 알려진 사실이다. 그러므로 최우 추정치와 같이 추정치가 제한조건을 만족시키지 못하는 것을 방지할 수 있다. 만약 사전정보가 없다면 무정보 분포에 제한조건만을 포함시킨 사전분포, 즉, $I_B(\beta)$를 사용할 수 있다. σ^2의 사전분포로는 역감마 분포 $IG(a,b)$혹은 무정보 사전분포 $\pi(\sigma^2) = \sigma^{-2}$를 사용할 수 있다.

3. 깁스표본기법

깁스 표본기법은 Gelfand and Smith (1990)가 제안한 이래 많은 응용분야에서 복잡한 문제들을 해결하는 도구로 사용되어 왔다. 깁스 표본기법의 특징은 모수벡터를 원소 모수로 분해하여 각 원소모수의 조건부 분포로부터의 난수 생성을 반복하면 일정 시간 후에 생성되는 모수벡터의 표본은 결합 사후분포를 따르게 된다는 것이다. 깁스 표본기법의 적용은 특히 모수에 대한 제한조건이 있을 때 유용하다. 비록 다차원 공간에서 복잡한 제한조건이라 할지라도 일차원 조건부 분포에서는 제한조건이 일차원 구간으로 표시되므로 이러한 제한조건을 만족시키는 표본의 생성이 매우 쉽기 때문이다.

부등 제한이 있는 선형회귀 모형의 경우 구체적으로 깁스 표본 알고리즘을 제시하면 다음과 같다. β의 사전분포가 $N(\mu_0, \Sigma_0)I_B(\beta)$로 주어지는 경우 β의 조건부 사후분포는

$$\begin{aligned}
\pi(\beta|\sigma^2, data) &\propto exp[-\frac{1}{2\sigma^2}(y - X\beta)^t(y - X\beta) - \frac{1}{2}(\beta - \mu_0)^t\Sigma_0^{-1}(\beta - \mu_0)] \\
&\propto exp[-\frac{1}{2}\{\beta^t(\frac{1}{\sigma^2}X^tX)\beta - 2\frac{1}{\sigma^2}\beta^tX^ty + \beta^t\Sigma_0^{-1}\beta - 2\beta^t\Sigma_0^{-1}\mu_0\}] \\
&\propto exp[-\frac{1}{2}\{\beta^t(\frac{1}{\sigma^2}X^tX + \Sigma_0^{-1})\beta - 2\beta^t(\frac{1}{\sigma^2}X^ty + \Sigma_0^{-1}\mu_0)\}] \\
&\propto pdf\ of\ N((\frac{1}{\sigma^2}X^tX + \Sigma_0^{-1})^{-1}(\frac{1}{\sigma^2}X^ty + \Sigma_0^{-1}\mu_0), \frac{1}{\sigma^2}X^tX + \Sigma_0^{-1})
\end{aligned}$$

이므로

$$\beta|\sigma^2, data \sim N(\mu^\pi, \Sigma^\pi)I_B(\beta),$$

$$\begin{aligned}
\Sigma^\pi &= (\sigma^{-2}X^tX + \Sigma_0^{-1})^{-1} \\
\mu^\pi &= \Sigma^\pi(\sigma^{-2}X^ty + \Sigma_0^{-1}\mu_0)
\end{aligned}$$

이며 무정보 사전분포 $I_B(\beta)$를 사용한 경우에는 위 식에서 $\Sigma_0^{-1} = 0$으로 두면 된다. σ^2의

사전분포로 $IG(a,b)$를 사용한 경우 σ^2의 사후분포는

$$\begin{aligned}\pi(\sigma^2|\beta,data) &\propto (\sigma^2)^{-n/2}exp[-\frac{1}{2\sigma^2}(y-X\beta)^t(y-X\beta)](\sigma^2)^{-(a+1)}exp[-\frac{b}{\sigma^2}] \\ &\propto (\sigma^2)^{-(n/2+a+1)}exp[-\frac{1}{\sigma^2}\{\frac{1}{2}(y-X\beta)^t(y-X\beta)+b\}] \quad (3.1) \\ &\propto pdf\ of\ IG(n/2+a,(y-X\beta)^t(y-X\beta)/2+b) \quad (3.2)\end{aligned}$$

이므로

$$\sigma^2|\beta,data \sim IG(n/2+a,(y-X\beta)^t(y-X\beta)/2+b)$$

이고 무정보 사전분포를 사용한 경우에는 위 분포에서 $a=0$, $b=0$을 대입하면 된다. 위의 사후분포를 보면 σ^2의 조건부 사후분포로부터 표본을 생성하기는 매우 쉽다.

제한조건 B를 만족시키는 사후분포로부터 β를 생성하는 가장 쉬운 방법으로는 기각기법이 있다. 이는 $N(\mu^\pi,\Sigma^\pi)$분포에서 β를 생성한 후 B를 만족하는 표본만 택하는 것이다. 그러나 기각기법은 제한조건 B를 만족하는 β의 확률이 작은 경우 매우 효율이 떨어지는 단점이 있다. 특히 다차원 공간에서는 대개 B의 확률이 극히 작은 경우가 발생하는데 이러한 경우 기각기법은 매우 비효율적일 수 있다.

B의 확률이 작을 경우 효율적으로 사용할 수 있는 기법으로는 역누적함수 기법이 있다. B_j를 $\{\beta_k, k\neq j\}$가 주어진 경우 B를 만족하는 β_j의 구간이라 하면 β_j의 조건부 사후분포는 $N(\mu_j^\pi,\Sigma_{jj}^\pi)\ I_{B_j}(\beta_j)$를 따르는데 여기서 μ_j^π, Σ_{jj}^π는 각각 $N(\mu^\pi,\Sigma^\pi)$에서 j 번째 원소의 조건부 평균과 분산을 나타낸다. B_j는 일차원 제한조건이므로 이를 서로 겹치지 않는 구간 $U_{jl}, l=1,...,m_j$, 의 합집합으로 표시할 수 있고 따라서 역누적함수 기법의 사용이 용이하게 이루어질 수 있다. 예를 들어 $m_j=1$ 인 경우, 즉, B_j가 하나의 구간 $U_j=(u_{1j},u_{2j})$으로 표시되는 경우에 역누적함수 기법을 사용하여 β_j의 사후 표본을 생성하는 알고리즘은 다음과 같다. $\Phi(\cdot)$를 표준정규분포의 누적분포함수라고 하면 균일 분포 (0, 1)을 따르는 난수 W와 β_j간에는

$$W = (\Phi(\frac{\beta_j-\mu_j^\pi}{\Sigma_{jj}^\pi}) - \Phi(\frac{u_{1j}-\mu_j^\pi}{\Sigma_{jj}^\pi}))/(\Phi(\frac{u_{2j}-\mu_j^\pi}{\Sigma_{jj}^\pi}) - \Phi(\frac{u_{1j}-\mu_j^\pi}{\Sigma_{jj}^\pi}))$$

의 관계식이 성립한다. 이 식을 β_j에 관해 풀면.

$$\beta_j = \Phi^{-1}[W\ (\Phi(\frac{u_{2j}-\mu_j^\pi}{\Sigma_{jj}^\pi}) - \Phi(\frac{u_{1j}-\mu_j^\pi}{\Sigma_{jj}^\pi})) + \Phi(\frac{u_{1j}-\mu_j^\pi}{\Sigma_{jj}^\pi})]\Sigma_{jj}^\pi + \mu_j^\pi$$

이다. 따라서 W를 균일분포로 부터 생성한 후 위 관계식을 이용하여 β_j로 변환하면 원하는 제한조건을 만족하는 β_j의 사후표본을 얻을 수 있다. 이 역누적함수 기법은 간단하고 효율적으로 원하는 제한 정규 분포로부터 표본을 생성하는 장점이 있다.

위와 같이 조건부 사후분포로부터 차례로 표본을 생성하는 과정을 반복하면 (β,σ^2)의 표본의 분포는 이들의 실제 결합사후분포로 수렴한다는 사실이 알려져 있다. 따라서 충분히 오랜 동안 깁스표본기법을 수행한 후 (β,σ^2)의 포본을 얻을 수 있다. 이처럼 일반적으로 어려운 제한조건이 있는 다차원 정규분포로부터의 표본생성을 깁스표본기법은 손쉽게 해결한다.

4. 베이지안 변수선택

변수 선택은 모형선택의 일종으로 볼 수 있다. 따라서 다음과 같이 2^p개의 모형을 고려한다. 먼저, 모형 M_0는 p개의 변수가 모두 포함된 전체모형 (full model) 이라 정의하고, $i = 1, ..., 2^p - 1$에 대하여 모형 M_i는 β의 i번째 분분집합 $\beta_{(i)}$가 0인 모형이라 정의하자. 그리고 $\beta_{-(i)}$는 β중 $\beta_{(i)}$를 제외한 나머지 원소들의 집합이라 정의한다.

베이지안 모형선택은 각 모형의 사후확률,

$$\pi(M_i|data) = \frac{\int p(data|\beta_{-(i)}, M_i)\pi(\beta_{-(i)}|M_i)d\beta_{-(i)}p(M_i)}{\sum_{i=0}^{2^p-1}\int p(data|\beta_{-(i)}, M_i)\pi(\beta_{-(i)}|M_i)d\beta_{-(i)}p(M_i)},$$

에 근거한다. 여기에서 $p(M_i)$는 모형 M_i의 사전확률을, $\pi(\beta_{-(i)}|M_i)$는 모형 M_i하에서 $\beta_{-(i)}$의 사전분포를 나타낸다.

만약 $\pi(\beta_{-(i)}|M_i) = \pi(\beta_{-(i)}|\beta_{(i)} = 0, M_0)$가 성립하고 $p(M_i)$가 모두 같으면,

$$\begin{aligned}
\frac{\pi(M_i|data)}{\pi(M_0|data)} &\propto \int p(data|\beta_{-(i)}, M_i)\pi(\beta_{-(i)}|M_i)d\beta_{-(i)} \\
&= \int p(data|\beta_{-(i)}, \beta_{(i)} = 0, M_0)\pi(\beta_{-(i)}|\beta_{(i)} = 0, M_0)d\beta_{-(i)} \\
&= \frac{\int p(data|\beta_{-(i)}, \beta_{(i)} = 0, M_0)\pi(\beta_{-(i)}, \beta_{(i)} = 0, M_0)d\beta_{-(i)}}{\pi(\beta_{(i)} = 0|M_0)} \\
&= \pi(\beta_{(i)} = 0|data, M_0)/\pi(\beta_{(i)} = 0|M_0) \\
&\equiv BF_i \qquad (4.1)
\end{aligned}$$

이 성립하는데 여기에서 $\pi(\beta_{(i)} = 0|data, M_0)$는 모형 M_0하에서 $\beta_{(i)}$의 주변 사후밀도함수의 0에서의 값이다. 따라서 각 모형의 사후확률은

$$\begin{aligned}
P(M_0|data) &= (1 + \sum_{i=1}^{2^p-1} BF_i)^{-1}, \\
P(M_i|data) &= BF_i \cdot P(M_0|data),\ i = 1, .., 2^p - 1. \qquad (4.2)
\end{aligned}$$

로 주어진다. 위의 사후확률이 주어지면 가장 큰 사후확률을 갖는 모형을 최적모형으로 선택하거나, 만약 몇 모형의 사후확률이 비슷하다면 그러한 모형들을 동시에 고려할 수 있겠다 (Hoeting et. al., 1999).

식 (4.1)을 보면 변수선택의 문제가 $2^p - 1$개의 주변 사후밀도함수의 계산문제로 귀착되었음을 알 수 있다. Chib (1995) 와 Chen (1994) 는 MCMC 표본으로부터 사후밀도함수를 추정하는 기법을 제안하였다. 그러나 이 방법들은 각 주변사후밀도함수 마다 새로운 MCMC 기법을 적용시켜야 하기 때문에 $2^p - 1$번의 MCMC 알고리즘의 수행이 요구되고 따라서 우리의 문제에는 적합치 않다. 반면에, Oh (1999)는 하나의 전체모형에 대한 MCMC 표본으로부터 모든 가능한 주변 사후밀도함수를 동시에 추정하는 기법을 제안하였는데 본 논문에서 다루는 문제와 같이 하나의 전체모형이 있고 그 안에 내포된 많은 모형들이 존재

할 경우 매우 적절한 방법이다. 더욱이 본 논문에서 다루는 문제와 같이 각 원소모수의 조건부 사후밀도함수들이 모두 주어진 경우에 적용이 매우 용이하다는 장점이 있다.

예시를 위하여 $\pi(\beta_1 = \beta_2 = 0|data, M_0)$와 $\pi(\beta_1 = \beta_2 = \beta_3 = 0|data, M_0)$를 추정하는 Oh(1999)의 알고리즘을 살펴보면,

$$\begin{aligned}\pi(\beta_1 = \beta_2 = 0|data, M_0) &= E[\pi(\beta_2 = 0|\beta_1 = 0, \beta_3, .., \beta_p, \sigma^2, data, M_0) \\ &\quad \times \pi(\beta_1 = 0|\beta_2, \beta_3, .., \beta_p, \sigma^2, data, M_0)]\end{aligned}$$

이고

$$\begin{aligned}\pi(\beta_1 = \beta_2 = \beta_3 = 0|data, M_0) &= E[\pi(\beta_3 = 0|\beta_1 = \beta_2 = 0, \beta_4, .., \beta_p, \sigma^2, data, M_0) \\ &\times \pi(\beta_2 = 0|\beta_1 = 0, \beta_3, .., \beta_p, \sigma^2, data, M_0) \\ &\times \pi(\beta_1 = 0|\beta_2, .., \beta_p, \sigma^2, data, M_0)]\end{aligned}$$

으로 나타낼 수 있는데 이 때 기대치는 (β, σ^2)의 결합사후분포에 대한 것이다. 따라서 전체 모형 M_0에 대하여 깁스표본기법을 적용시킨 후 얻어진 표본을 사용하여 필요한 기대치를 모두 추정할 수 있고 따라서 임의의 주변 사후 밀도함수 값을 동시에 계산할 수 있다. 이 주변 사후 밀도함수 값과 사전밀도함수 값으로부터 각 모형의 사후확률을 식 (4.2)와 같이 계산하여 모형선택, 즉, 변수선택에 활용한다.

5. 예: Michigan 대학생들의 임대료

Pindyck 과 Rubinfeld (1998)는 Michigan 대학의 학생들을 대상으로 임대료, 임대한 방의 갯수, 거주자 수, 성별, 학교에서 집까지의 거리에 관한 32개의 관측치들을 제시하고, 임대료를 종속변수로 두고 나머지 변수를 독립변수로 두어 선형모형을 추정하였다.

개인당 임대료를 y_i, 개인당 방의 수를 r_i, 캠퍼스에서의 거리를 d_i, 성별을 지시변수 s_i (남자면 1, 여자면 0) 로 두고 다음과 같은 선형 모형을 가정한다.

$$y_i = \beta_1 + \beta_2 s_i r_i + \beta_3(1 - s_i)r_i + \beta_4 s_i d_i + \beta_5(1 - s_i)d_i + \epsilon_i,$$

$$\epsilon \sim N(0, \sigma^2 I).$$

이 모형에서 σ^2은 오차분산, $\beta_i, i = 1, .., 5$는 회귀계수로, σ^2과 $\beta = (\beta_1, ..., \beta_5)$가 알려지지 않은 모수가 된다. 그런데 사용하는 방의 수가 많을수록 임대료가 비싸고 학교에서의 거리가 멀수록 임대료가 저렴할 것으로 예상되므로 $\beta_2 \geq 0$, $\beta_3 \geq 0$, $\beta_4 \leq 0$, $\beta_5 \leq 0$의 부호가 예상되어 이를 β_i에 대한 제한조건으로 두면 $\beta = (\beta_1, .., \beta_5)^t$의 제한공간 B는

$$B = \{\beta\ ;\ -\infty \leq \beta_1 \leq \infty,\ \beta_2 \geq 0,\ \beta_3 \geq 0,\ \beta_4 \leq 0,\ \beta_5 \leq 0\}$$

이 된다 (Geweke, 1986).

β의 사전분포로는 제한이 있는 다변량 정규분포 $N(\mu_0, \Sigma_0)I_B(\beta)$ 로 두고 μ_0는 보통 최소제곱 추정치(Ordinary Least Square Estimator) $\hat{\beta}^O = (X^tX)^{-1}X^ty$로 두고 Σ_0는 우도함수

	Bayes		OLS	
	평균	표준편차	평균	표준편차
β_1	37.401	33.488	38.258	32.256
β_2	136.872	36.936	103.860	38.494
β_3	123.779	38.510	122.340	37.385
β_4	-0.847	0.793	3.314	1.960
β_5	-1.183	0.549	-1.154	0.571

표 5.1: 베이즈 추정치와 보통 최소제곱 추정치의 비교

로부터 β의 분산이 $\sigma^2(X^tX)^{-1}$인 점을 감안하여 사전정보가 대략 한 개의 관측치가 가지는 정보에 해당하도록 (unit information prior) Σ_0를 $\Sigma_0 = n\sigma^2(X^tX)^{-1}$로 둔다. 다음, σ^2의 사전분포로는$IG(3, 2s^2)$을 선택하여 사전평균이 s^2이고 분산이 큰 분포가 되도록 하는데 이 때 s^2은 보통 최소제곱 추정치에 해당하는 잔차평균으로 $s^2 = (y - X\hat{\beta}^O)^t(y - X\hat{\beta}^O)/(n-p)$이다.

3절에서 기술된 깁스표본기법을 적용하여 $\sigma^2, \beta_1, ..., \beta_5$ 각각의 조건부 분포로부터 표본을 반복적으로 생성하는데, β 의 초기값 $\beta^{(0)}$을 보통 최소제곱 추정치 $\hat{\beta}^O$로 주고 σ^2의 초기값은 $\sigma^{2(0)} = s^2$으로 준다. 이 초기치를 사용하여 깁스표본기법을 25000번 수행한 다음 처음 5000번 이후에 생성된 20000개의 표본으로부터 각 모수의 추정치와 표준오차를 구하고 이를 보통 최소제곱 추정치(OLS estimate) 와 비교한 결과는 표 5.1 과 같다. 참고로, 보통 최소제곱 추정치에 대한 표준편차는 $\sqrt{s^2(X^tX)^{-1}}$의 대각원소의 값이다.

위 결과를 보면 β_4를 제외한 모수들은 보통 최소제곱 추정치와 비슷한 값으로 추정되었으나, β_4의 경우 깁스표본기법에 의한 추정치는 음수이나 보통 최소제곱 추정치는 양수로 상당히 큰 차이가 있다. 학교에서의 거리가 멀수록 임대료는 저렴할 것으로 기대되므로 보통 최소제곱 추정치 보다는 음수의 결과를 보인 베이지안 추정치가 더 좋은 결과임을 알 수 있다. 또한 β_4의 표준편차를 보면 베이즈 추정의 표준편차가 보통 최소제곱 추정의 표준편차 보다 현저하게 작은데 이는 보통 최소제곱 추정에서는 모든 실수구간을 고려한데 반하여 베이즈 추정에서는 음수로 제한된 구간을 고려하였기 때문이다.

깁스표본기법에 의한 모의실험에서 생성된 표본들의 사후주변밀도함수는 그림 5.1과 같다. 그림에서 보면 β_i는 β_4를 제외하고 모두 정규분포의 형태를 따르고 있으며 보통 최소제곱 추정치와 베이지안 추정치가 크게 차이가 나는 β_4는 0에서 잘려진 정규분포 (truncated normal distribution)의 형태를 따르고 있음이 현저하게 드러난다.

다음으로 모형선택, 즉, 변수선택의 문제를 살펴보도록 하겠다. 본 예제에서는 모형 M_0하에서 $2^5 - 1 = 31$ 개의 부분집합에 대하여 $\beta_{(i)} = 0$일 때의 $\beta_{(i)}$의 사후주변밀도함수의 값, $\pi(\beta_{(i)} = 0|data, M_0)$의 계산이 필요하다. 이를 계산한 결과가 표 5.2 이고 이로부터 식 (4.2)를 사용하여 가능한 모형의 사후확률을 계산한 결과 다음 5 개의 모형이 유의한 사후확률을 갖는다 (표 5.3). 따라서 가장 유의한 모형은 β_1과 β_4가 빠진 모형이고 나머지 4개의 모형들이 어느 정도 자료를 설명하는 측면이 있다고 보여진다. 이 결과는 표 5.1

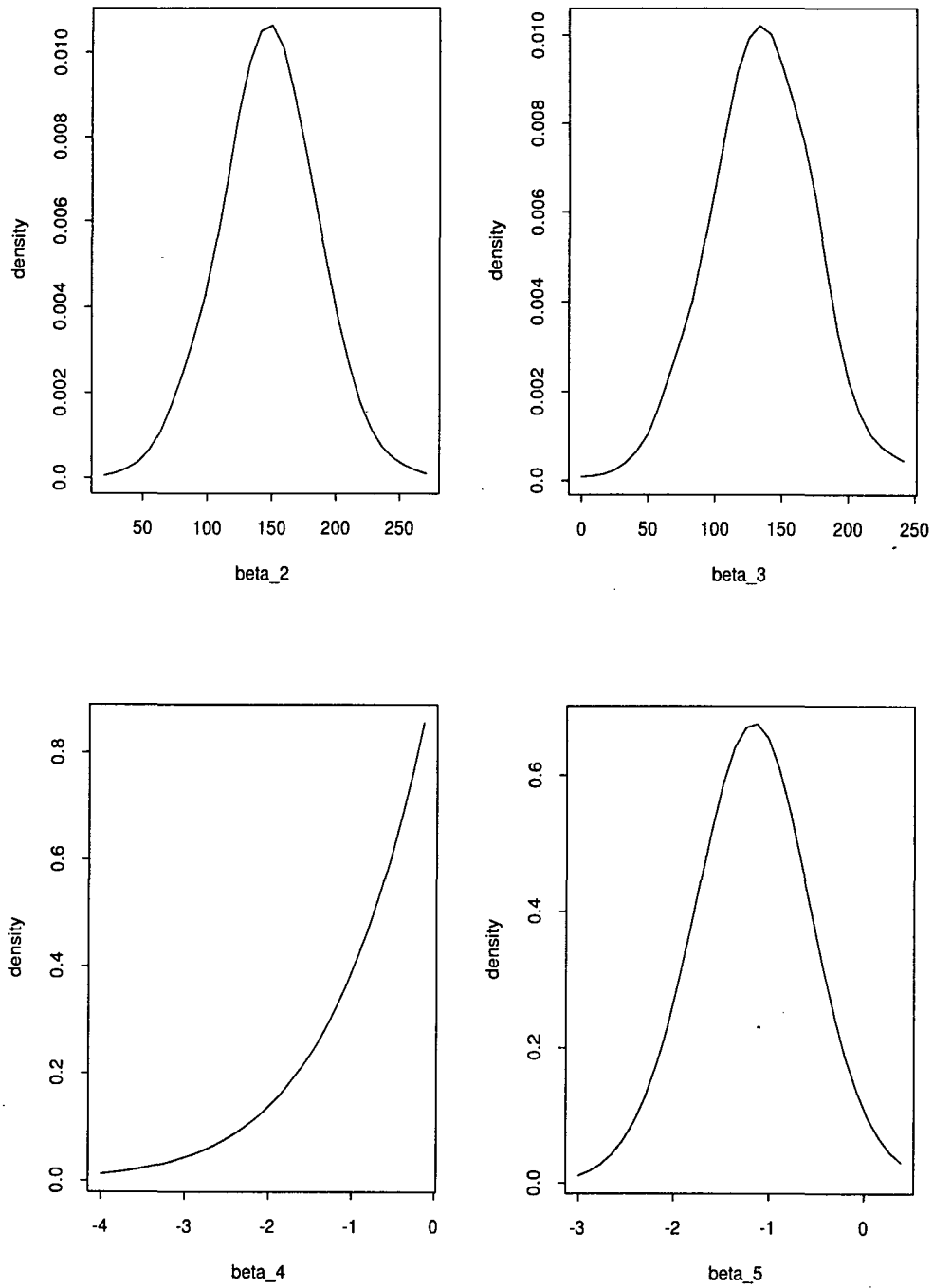

그림 5.1: β 의 주변사후밀도함수

$\beta_{(i)}$	$\pi(\beta_{(i)}=0\|data)$	$\beta_{(i)}$	$\pi(\beta_{(i)}=0\|data)$
β_1	6.544E-03	β_1,β_2,β_3	6.407E-51
β_2	4.684E-06	β_1,β_2,β_4	2.314E-24
β_3	4.636E-05	β_1,β_2,β_5	5.627E-25
β_4	1.059E 00	β_1,β_3,β_4	9.225E-33
β_5	1.018E-01	β_1,β_3,β_5	3.925E-31
β_1,β_2	5.870E-25	β_1,β_4,β_5	6.527E-04
β_1,β_3	3.269E-32	β_2,β_3,β_4	2.161E-07
β_1,β_4	6.891E-03	β_2,β_3,β_4	1.715E-08
β_1,β_5	6.662E-04	β_2,β_4,β_5	5.100E-07
hline β_2,β_3	1.349E-07	β_3,β_4,β_5	1.548E-07
β_2,β_4	8.402E-06	$\beta_1,\beta_2,\beta_3,\beta_4$	2.532E-50
β_2,β_5	2.947E-07	$\beta_1,\beta_2,\beta_3,\beta_5$	7.669E-50
β_3,β_4	4.171E-05	$\beta_1,\beta_2,\beta_4,\beta_5$	2.212E-24
β_3,β_5	1.971E-05	$\beta_1,\beta_3,\beta_4,\beta_5$	1.111E-31
β_4,β_5	1.011E-01	$\beta_2,\beta_3,\beta_4,\beta_5$	2.834E-08
		$\beta_1,\beta_2,\beta_3,\beta_4,\beta_5$	3.030E-49

표 5.2: 사후확률밀도함수 값

Model	Prob
$\beta_1 = 0$	0.0385
$\beta_4 = 0$	0.1922
$\beta_1 = \beta_4 = 0$	0.3905
$\beta_4 = \beta_5 = 0$	0.1018
$\beta_1 = \beta_4 = \beta_5 = 0$	0.1975

표 5.3: 모형의 사후확률

의 결과와 일치하는 면이 있다. 표 5.1에서 보면 β_1과 β_4가 0과 유의하게 다르지 않고 β_5는 그다지 유의하지 않은데 표 5.3의 모형선택이 이를 잘 반영하고 있다.

6. 요약 및 제언

본 논문에서는 부등 제한조건이 있는 정규 선형 모형에서 몬테칼로 기법을 통한 베이지안 변수선택 방법을 제안하였다. 변수선택은 모형선택의 일종으로 각 모형의 사후확률 계산이 요구된다. 모든 변수를 포함하는 전체 모형과 변수들의 부분집합만을 포함하는 내포된 모형 간에 사후확률의 비(ratio)인 베이즈 상수는 부분 모수의 사후 밀도함수와 사전 밀도함수의 비로 표시될 수 있다. 보통 사전 밀도함수는 주어지나 사후 밀도함수는 부등 제한조건 때문에 수리적으로 주어지지 않는다. 이 문제를 본 논문에서는 Oh(1999)의 기법을 사용하여 해결한다. 실제 자료에 본 기법을 적용한 결과 합리적인 결론에 도달함을 알 수 있었다.

정규 선형모형 이외에 로짓이나 로그선형 등 일반 선형모형에도 본 기법의 적용을 고려해 볼 수 있다. 그러나 정규선형이 아닌 일반선형 모형에서는 계수 β_i의 조건부 사후분포가 편리한 형태로 주어지지 않아 기법의 직접적인 적용이 불가능하다. 이에 대한 연구는 추후 연구과제로 남겨 놓기로 한다.

참고문헌

[1] Chen, M-H. (1994), Importance-Weighted Marginal Bayesian Posterior Density estimation, *Journal of the American Statistical Association*, vol. 89, 818-824.

[2] Chib, S. (1995), Marginal Likelihood from the Gibbs Output, *Journal of the American Statistical Association*, vol. 90, 1313-1321.

[3] 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.

[4] Geweke, J. (1986), Exact Inference in the Inequality Constrained Normal Linear Regeression Model, *Journal of Applied Econometrics*, vol. 1, 127-141

[5] Geweke, J (1993), Bayesian Inference for Linear models Subject to Linear Inequality Constraints, Technical Report, Dept. of Econometrics, Univ. of Minesota

[6] Hoeting, J.A., Madigan, D., Raftery, A.E., and Volinsky, C.T. (1999), Bayesian Model Averaging: A Tutorial, *Statistical Science*, vol. 14, 382-412.

[7] Oh, M-S. (1999). Estimation of Posterior Density Functions from a Posterior Sample, *Computational Statistics and Data Analysis*, vol. 29, 411-427.

[8] Pyndyck, R.S. and Rubinfeld, D.L. (1998), *Econometric Models and Econometric Forecasts*, 3rd ed., McGraw-Hill, New-York.

[9] Raftery, A. E. (1996), *Hypothesis testing and model selection*, in: Gilks, W.R., Richardson, S., and Spiegelhalter, D.J. (Eds.), Markov chain Monte Carlo in practice, Chapman & Hall.

[2001년 8월 접수, 2002년 12월 채택]

Bayesian Variable Selection in Linear Regression Models with Inequality Constraints on the Coefficients

Man-Suk Oh [1)]

ABSTRACT

Linear regression models with inequality constraints on the coefficients are frequently used in economic models due to sign or order constraints on the coefficients. In this paper, we propose a Bayesian approach to selecting significant explanatory variables in linear regression models with inequality constraints on the coefficients. Bayesian variable selection requires computation of posterior probability of each candidate model. We propose a method which computes all the necessary posterior model probabilities simultaneously. In specific, we obtain posterior samples form the most general model via Gibbs sampling algorithm (Gelfand and Smith, 1990) and compute the posterior probabilities by using the samples. A real example is given to illustrate the method.

Keywords: Normal linear regression model, posterior probability, Markov chain Monte Carlo.

1) Associate Professor, Dept. of Statistics, Ewha Women's University, Sodaemoon Gu, Seoul 120-750. KOREA

E-mail: msoh@mm.ewha.ac.kr