한국통계학회논문집 제4권 2호
1997년 8월 pp. 333 - 343

깁스표본기법을 이용한 설명변수 선택문제에서 사전분포의 설정
- 선형회귀모형을 중심으로 -

박 종 선[1], 남궁 평[2], 한 숙 영[3]

요 약

선형회귀분석에서 변수의 선택문제는 최적의 모형을 찾는데 아주 중요한 부분을 차지한다. George와 McCulloch(1993)는 계층적 베이즈 모형과 깁스표본법을 이용하여 선형회귀모형에서 변수를 선택하는 문제를 고려하였다. 이 논문에서는 George와 McCulloch의 모형을 바탕으로 각각의 설명변수가 모형에 포함될 사전확률을 객관적인 기준에 의하여 결정하는 문제를 고려하여 보았다.

1. 서론

선형회귀분석은 일반적으로 p개의 설명변수 $X_1, X_2, \cdots, X_p$를 이용하여 하나의 종속변수 Y를 설명변수들의 선형결합으로 설명하려는 시도로서 이 때 변수선택의 문제는 p개의 설명변수에서 추출된 q $(q \leq p)$개의 $X_1^*, X_2^*, \cdots, X_q^*$로 이루어지는 최적의 모형

$$Y = \beta_0 + \beta_1 X_1^* + \beta_2 X_2^* + \cdots, + \beta_q X_q^* + \varepsilon$$

을 찾는데 그 목적이 있다. 이러한 변수선택의 문제는 최근까지 여러 방법이 연구되어 왔는데 대표적인 것으로는 C_p 방법(Mallows, 1973), Akaike의 정보규준(Akaike's Information Criterion - AIC, Akaike, 1973), 재표본을 이용한 방법(resampling methods), 깁스 표본기법을 이용한 방법 등이 있다. 이 중에 1993년 George와 McCulloch는 계층적 베이즈(Hierarchical Bayes)모형과 깁스(Gibbs) 표본기법을 이용하여 변수를 선택하는 문제를 고안하였다. 다변량분포에서 표본을 추출하는 한 방법인 깁스표본기법은 형상복원(image restoration) 등에 적용되어 왔으며, 90년대에 들어 비로소 Gelfand와 Smith (1990)에 의하여 통계분야에 적용이 알려지기 시작하면서 각광을 받고 있다. 이 논문에서는 George와 McCulloch의 모형에서 회귀계수가 모형에 포함될 사전확률을 모든 설명변수를 포함하는 모형 (full model)을 최소제곱법(least squares method)으로 적합하였을 때 각 회귀계수의 t값에 따라 객관적으로 부여함으로서 변수의 선택에 나타나는 효과를 살펴보았다.

1) (110-745) 서울 특별시 종로구 명륜동 3가 53, 성균관대학교 경상대학 통계학과 조교수
2) (110-745) 서울 특별시 종로구 명륜동 3가 53, 성균관대학교 경상대학 통계학과 교수
3) (110-745) 서울 특별시 종로구 명륜동 3가 53, 성균관대학교 대학원 경제학 석사

2. 깁스표본을 이용한 변수선택

George와 McCulloch(1993)는 회귀분석문제에서 주어진 설명변수 $X_1, \cdots, X_p$ 에서 모형에 필요한 부분집합들을 선택하기 위해 SSVS(Stochastic Search Variable Selection)라 부르는 절차를 발전시켰다. SSVS에서는 회귀계수들이 0인 경우와 아닌 경우에 따라 분산이 다른 상이한 분포를 가정하였다. 각각의 회귀계수가 모형에 포함될 확률을 베르누이 분포에서의 성공확률로 가정한 계층적 베이즈모형(Hierarchical Bayes Model)을 설정하였으며 최적의 모형선택을 위하여 사후분포로부터 깁스표본기법(Gibbs sampling technique)을 이용하여 추출된 모형중에서 출현빈도가 높은 모형들을 살펴보게 된다.

2.1 계층적(Hierarchical) 베이즈 모형

다음과 같은 회귀모형을 가정하자.

$$Y \mid \beta, \sigma^2 \sim N_n(X\beta, \sigma^2 I) \tag{1}$$

여기서 반응변수 Y는 $n \times 1$ 벡터이고, 설명변수 $X=(X_1, \cdots, X_p)$ 는 $n \times p$ 인 행렬로 가정하자. $\beta=(\beta_1, \cdots, \beta_p)$ 와 σ^2 은 각각, 회귀계수와 분산항이다. 변수 선택과 관련된 정보를 알기 위해서 George와 McCulloch는 계층적 모형의 한 부분으로써 (1)을 고려했다. 여기서는 각각의 β_i 가 모형에 필요한 지 아닌 지에 따라 잠재(latent)변수 $\gamma_i = 1$ 또는 $\gamma_i = 0$의 값을 주어 정규혼합(normal mixture)모형을 다음과 같이 가정하였다.

$$\beta_i \mid \gamma_i \sim (1-\gamma_i)\, N(0, \tau_i^2) + \gamma_i N(0, c_i^2 \tau_i^2). \tag{2}$$

이 때

$$P(\gamma_i=1) = 1 - P(\gamma_i=0) = p_i \tag{3}$$

로 가정하였으며 여기서 p_i는 i번 째의 회귀계수가 0이 아닐 확률을 의미한다. c_i와 τ_i는 적절한 상수로, 특히, $\gamma_i=1$일 때 c_i(항상 $c_i>1$)가 큰 값을 갖도록하여 β_i가 0주위에서 넓게 퍼져 해당되는 X_i가 모형에 필요하도록 만들었다.

$\beta_i \mid \gamma_i$의 사전분포 (2)를 결합하면 다음과 같은 다변량 정규분포가 된다.

$$\beta \mid \gamma \sim N_p(0, D_\gamma R D_\gamma). \tag{4}$$

여기서 $\gamma=(\gamma_1, \gamma_2, \cdots, \gamma_p)$이고, R 은 β 에 대한 사전 분포 상관 행렬로 생각할 수 있으며

$$D_\gamma \equiv diag\,[a_1\tau_1, \cdots, a_p\tau_p] \tag{5}$$

이고 이 때 $\gamma_i=0$ 이면 $a_i=1$ 이고, $\gamma_i=1$ 이면 $a_i=c_i$가 되어 (2)와 같아진다.

베르누이 모형(3)은 γ의 값이 2^p개의 가능한 조합을 갖는 이산형 분포 $f(\gamma)$의 주변분포로서 사용되었다. 변수선택의 목적을 제외하고, $f(\gamma)$는 γ가 정확하게 $r_i=1$이 되는 사전확률인데 이러한 β_i는 모형에서 0이 아닌 추정값을 갖게 된다.

계층적 모형의 마지막 구성 요소는 잔차분산 σ^2의 사전분포이며 여기서는 역 감마 공액 사전(inverse gamma conjugate priors)분포를 사용한다.
따라서

$$\sigma^2 \mid \gamma \sim IG(\frac{\nu_\gamma}{2}, \frac{\nu_\gamma \lambda_\gamma}{2}) \tag{6}$$

또는

$$\frac{\nu_\gamma \lambda_\gamma}{\sigma^2} \sim X^2_{\nu_\gamma}$$

을 가정하였다. 여기서 ν_γ는 자유도이다.

$f(\gamma)$의 선택에서는 $X_1, \cdots, X_p$ 중 부분 집합에 포함되는 X들을 위하여 사전 정보를 결합해야 한다. 특히, p가 큰 경우에 2^p 개 선택이 가능하여 p가 크면 그 수가 기하급수적으로 커지게 된다. 주변 분포 (3)에서 γ_i 들이 독립인 경우를 생각할 수 있는데 이 때 $f(\gamma)$의 사전분포는

$$f(\gamma) = \prod_{i=1}^{p} p_i^{r_i}(1-p_i)^{(1-r_i)} \tag{7}$$

이 된다.

이상을 종합하여 주어진 자료 Y에 따라 구해진 사후분포

$$f(\gamma \mid Y) \propto f(Y \mid \gamma) \times f(\gamma) \tag{8}$$

를 통해 우리는 최종 모형에 포함될 X_i들을 결정하게 되는데 이때 깁스표본기법을 이용하여 사후분포(8)로부터 γ의 표본들을 추출하여 결정한다.

2.2 깁스표본기법

깁스표본기법은 2차원 이상의 다변량 분포에서 직접 표본추출이 불가능한 경우에 조건부 분포들로부터 간접적으로 확률변수를 생성하는 방법으로서 Metropolis, Rosenbluth, Rosenbluth, Teller와 Teller에 의해서 1953년에 소개되었으며, Hastings (1970), Geman과 Geman (1984), Gelfand와 Smith (1990) 등에 의해서 발전되었다. 이 기법은 다른 기법으로는 해결이 곤란한 복잡한 함수의 적분에 쉽게 사용할 수 있고 특히, 통계분석에서 발생하는 다차원 적분에 유용하여 베이지안 통계, 변량이 있는 선형모형분석, 판별분석 등에 널리 사용되어 왔다. 그러나 효율이 낮고 추정값의 분산을 계산하기 어려우며 깁스표본기법의 적용을 위해서는 각 적분변수의 조건부 확률분포로부터 확률난수의 생성이 쉬운 형태이어야 한다는 조건이 필요하다.

2.3 최적모형의 선택

2.3.1 최적모형을 위한 깁스표본기법

최적모형을 결정하기 위하여 깁스표본기법을 사용하여 다음과 같은 "깁스 수열"을 생성하며 이를 위한 구체적인 단계들은 아래와 같다.

$$\beta^{(0)}, \sigma^{(0)}, \gamma^{(0)}, \beta^{(1)}, \sigma^{(1)}, \gamma^{(1)}, \cdots, \beta^{(j)}, \sigma^{(j)}, \gamma^{(j)}, \cdots, \tag{9}$$

1단계>> (1) 의 최소제곱 추정값을 이용하여 $\beta^{(0)}, \sigma^{(0)}$ 그리고 $\gamma^{(0)} \equiv (1, 1, \cdots, 1)$로 초기화한다.

2-1단계>> 벡터 $\beta^{(j)}$ 는 다음과 같은 사후분포에 의해서 얻어진다.

$$\begin{aligned} \beta^{(j)} &\sim f(\beta^{(j)} \mid Y,\ \sigma^{(j-1)}, \gamma^{(j-1)}) \\ &= N_p(A_{\gamma^{(j-1)}} (\sigma^{(j-1)})^{-2} X'X\, \hat{\beta}_{LS}, A_{\gamma^{(j-1)}}) \end{aligned} \tag{10}$$

여기서,

$$A_{\gamma^{(j-1)}} = ((\sigma^{(j-1)})^{-2} X'X + D^{-1}_{\gamma^{(j-1)}} R^{-1} D^{-1}_{\gamma^{(j-1)}})^{-1} \tag{10a}$$

$$D_\gamma^{-1} = diag[(a_1\tau_1)^{-1}, \cdots, (a_p\tau_p)^{-1}]$$

2-2단계>> 분산 $\sigma^{(j)}$ 은 다음과 같은 사후분포에 의해서 얻어진다.

$$\begin{aligned} \sigma^{(j)} &\sim f(\sigma^{(j)} \mid Y,\ \beta^{(j)}, \gamma^{(j-1)}) \\ &= IG\left(\frac{n + \nu_{\gamma^{(j-1)}}}{2}, \frac{\mid Y - X\beta^{(j)} \mid^2 + \nu_{\gamma^{(j-1)}} \lambda_{\gamma^{(j-1)}}}{2}\right) \end{aligned} \tag{11}$$

2-3단계>> 벡터 $\gamma^{(j)}$ 도 사후 조건부 분포로부터 얻어진다.

$$\gamma^{(j)}{}_i \sim f(\gamma^{(j)}{}_i \mid Y, \beta^{(j)}, \sigma^{(j)}, \gamma^{(j)}{}_{(i)}) = f(\gamma^{(j)}{}_i \mid \beta^{(j)}, \sigma^{(j)}, \gamma^{(j)}{}_{(i)}) \tag{12}$$

여기서,

$$\gamma^{(j)}{}_{(i)} = (\gamma^{(j)}{}_i, \cdots, \gamma^{(j)}{}_{i-1}, \gamma^{(j-1)}{}_{i+1}, \cdots, \gamma^{(j-1)}{}_p)$$

각각의 i에 대하여 분포(12)는 다음과 같은 베르누이 분포를 따르게 된다.

$$P(\gamma^{(j)}{}_i = 1 \mid \beta^{(j)}, \sigma^{(j)}, \gamma^{(j)}{}_{(i)}) = \frac{a}{a+b}. \tag{13}$$

여기서

$$a=f(\beta^{(j)}\mid \gamma^{(j)}_{(i)},\ \gamma^{(j)}_{i}=1)\ \times\ f(\sigma^{j}\mid \gamma^{(j)}_{(i)},\ \gamma^{(j)}_{i}=1)\ \times\ f(\gamma^{(j)}_{(i)},\ \gamma^{(j)}_{i}=1) \quad (14a)$$

이고

$$b=f(\beta^{(j)}\mid \gamma^{(j)}_{(i)},\ \gamma^{(j)}_{i}=0)\ \times\ f(\sigma^{j}\mid \gamma^{(j)}_{(i)},\ \gamma^{(j)}_{i}=0)\ \times\ f(\gamma^{(j)}_{(i)},\ \gamma^{(j)}_{i}=0) \quad (14b)$$

이다.

(10), (11), (12)으로부터 반복적으로 연속된 표본을 추출하면 깁스 수열(9)를 얻을 수 있다. 수열(9)는 기하학적으로 수렴하는 동질인 에르고딕 마코브 체인 (homogeneous ergodic Markov chain) 이다. 따라서 그것은 유일한 정상분포(stationary distribution) $f(\gamma\mid Y)$를 갖게 된다. 이 수열표본을 반복적(iteratively)으로 추출하게 되면, γ의 실제 값의 경험적인 분포는 실제 사후 분포 $f(\gamma\mid Y)$에 수렴할 것이다. 따라서 구해진 깁스수열 $\gamma^{(j)}$ 중에서 빈도가 높은 γ에 따른 모형이 최적일 가능성이 높아진다.

3. 사전분포 및 모수설정

3.1절에서는 먼저 George와 McCulloch가 설정한 파라메터들에 대하여 살펴보고 3.2절에서는 γ에 대한 사전분포로 모든 설명변수를 포함하는 모형에서 각 회귀계수들의 t값 또는 p-값에 따라 각 설명변수가 모형에 포함될 사전확률을 부여하는 방법에 대하여 살펴보자.

3.1 파라메터들의 설정

3.1.1 τ_i와 c_i의 선택

식 (2)에서 $\gamma_i=0$일 때 $\beta_i \sim N(0,\tau_i^2)$이므로 이때 β_i를 0으로 할 수 있도록 τ_i를 선택하여야 한다. 또한 $c_i\,(>1)$는 $\gamma_i=1$일 때 $\beta_i \sim N(0,c_i^2\tau_i^2)$이므로 β_i가 0이 아닌 값으로 추정될 수 있도록 큰 값으로 선택되어야 한다.

3.1.2 R의 선택

행렬 R은 γ 가 주어졌을 때 β의 사전분포 상관행렬로서 β의 사후분포 공분산 행렬은

$$(\sigma^{-2}X'X+D_\tau^{-1}R^{-1}D_\tau^{-1})^{-1} \quad (14)$$

이 되는데 극단적인 두 경우인 $R=I$ 그리고 $R\propto(X'X)^{-1}$를 보면 $R=I$ 일 때, β의 요소는 $\pi(\beta\mid\gamma)$에서 독립이고, $R\propto(X'X)^{-1}$일 때 사전 분포 상관은 설명변수의 상관계수(design correlation)와 같아진다.

3.1.3 ν_γ와 λ_γ의 선택

역 감마 사전 분포(inverse gamma prior) (6)에서 ν_γ와 λ_γ를 선택을 할 때 ν_γ는 관측값들의 수이고 $[\frac{\nu_\gamma}{(\nu_\gamma-2)}]\lambda_\gamma$은 σ^2의 사전 분포 추정값이라는 점을 고려할 수 있다.

3.2 γ의 사전분포

George와 McCulloch는 γ에 대한 사전분포로 무차별 사전분포 (indifference prior distribution) 즉, 각 γ_i의 사전확률인 p_i를 0.5로 주어 해당하는 설명변수 X_i가 모형에 포함될 확률을 0.5로 하는 경우만을 생각하였다. 그러나 이 경우에는 결과가 τ_i와 c_i의 값에 많은 영향을 받아 이 값들이 적당하지 않는 경우 합리적이지 못한 결과들이 나오는 것을 볼 수 있다. 여기서 우리는 γ에 대한 사전확률을 조정하는 한 방법으로서 모든 설명변수가 포함되는 모형 (full model)에 일반최소제곱법(ordinary least squares method)을 적용하여 나타난 각 회귀계수들의 t값이 각 설명변수들에 대한 모형에의 공헌도를 나타낸다는 가정하에 이를 이용하는 방법을 생각하였다.

일반적으로 회귀계수가 0인 지 아닌 지에 대한 가설검정에서는 각 계수의 t값에 따른 p-값의 크기에 따라 각 계수가 0인지 아닌 지에 대한 판단을 하게 된다. p-값은 0과 1사이의 실수값으로 각 계수에 대한 p-값을 직접 p_i라 놓아도 논리적으로 아무 문제도 발생하지 않는다. 그러나 나타난 p-값이 0 또는 1에 근접해 있을 경우 이에 해당되는 설명변수가 항상 모형에 포함되거나 또는 모형에 거의 포함되지 않는 경우가 발생할 수 있다. 이러한 문제점을 해결하기 위하여 나타난 p-값을 선형변환하여 0에서 1사이가 아닌 예를 들면 0.2에서 0.8또는 0.3에서 0.7사이의 값을 갖도록 조정하여 사용하였다. 특히, t값이나 p-값은 설명변수들이 독립이 아닐 경우에 왜곡되는 경향이 있으므로 모의실험에서는 독립인 설명변수와 그렇지 않은 경우(다중공선성이 있는 경우)에 대하여 모두 기존의 방법과 비교하였다. 모든 계산은 George와 McCulloch가 작성한 Fortran 코드를 Xlisp-Stat (Tierney, 1990)으로 변환하여 수행하였다.

4. 모의실험과 사례분석

4.1 모의실험

예제 4.1 이 예제는 2개의 표본으로 구성되었는 데 관측값의 수가 $n=100$이고, 독립 변수가 5 ($p=5$)개인 회귀모형을 생각하였다. 첫 번째 표본(표본1)에서는 예측변수 $X_1,\cdots,X_5$는 독립 표준정규벡터(independent standard normal vectors) 즉 $N_5(0,1)$에서 추출하였다. 종속변수 Y는

$$Y=X_1+1.5X_2+\varepsilon \quad , \quad \varepsilon \sim N_{100}(o,\sigma^2 I)$$

에 의하여 생성되었고, 계수 $\beta=(1,\ 1.5,\ 0,\ 0,\ 0)'$으로 σ는 0.5, 1.0, 1.5의 세가지 경우를 생각

하였다. 두 번째 표본(표본2)은 설명변수들이 서로 상관관계를 갖도록 X_3이 $X_3^* = X_1 + 0.1Z$ ($Z \sim N(0,1)$) 로 대체되었다는 사실만 다르고 모형은 전과 같다. 비록 회귀 계수의 값이 (표본1)과 같지만, 계수 추정값 $\hat{\beta}_3$ 와 $\hat{\beta}_1$는 공선성(collinearity)에 의해 그 값이 증가한다.

(표본1)과 (표본2)에서 공통적인 관심인, 즉 사전 분포를 $p_i = 0.5$인 무차별 사전분포를 준 경우와 p-값을 사용해서 $0.2 < p_i < 0.8$ 와 $0.3 < p_i < 0.7$ 인 경우 SSVS방법을 비교하기 위하여 γ에 대한 5000개의 깁스표본을 사후분포로부터 추출하고 그 결과를 표와 그림을 통해서 관찰하였다. 표 5.1 과 표 5.2 에서는 γ의 빈도가 가장 높은 모형 5개를 나타냈고, 변수들도 추출된 빈도의 크기에 따라 나열하였다. 이 때 σ는 결과에 많은 영향을 주지 않는데 반하여 사전 분포 p_i 값은 결과에 많은 영향을 주고 있음을 알 수 있다. 먼저 (표본1)의 경우 무차별 사전분포(indifference prior)인 $p_i = 0.5$ 보다 $0.2 < p_i < 0.8$ 이나 $0.3 < p_i < 0.7$ 과 같이 사전 분포를 준 경우가 더 좋은 결과를 주는 것을 볼 수 있다. (표본2)에서는 (표본1)보다 좋은 결과를 주지는 않지만 여기서도 역시 사전분포를 $0.2 < p_i < 0.8$ 이나 $0.3 < p_i < 0.7$를 준 경우가 무차별 사전분포보다 좋은 결과를 보여주고 있다(표 4.1 및 4.2 참조).

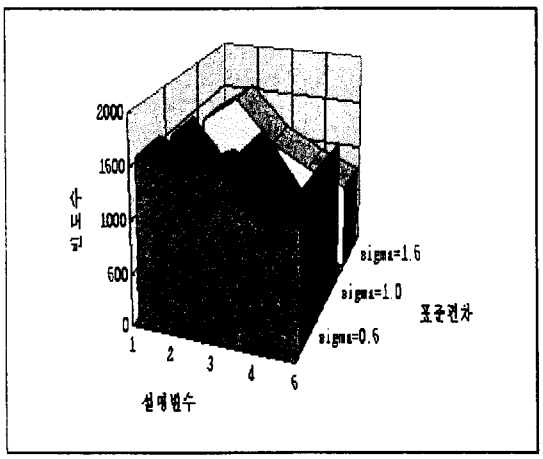
전체적으로 $0.2 < p_i < 0.8$이나 $0.3 < p_i < 0.7$을 사용할 경우 X_1, X_2가 포함된 최적모형의 빈도수가 모든 경우에 가장 높지는 않았지만 X_1과 X_2가 중요하다는 것은 잘 나타나 있다고 하겠다. $p_i = 0.5$를 사용할 경우는 모든 문제에서 설명변수를 하나만 포함하는 모형이 최적인 것으로 나타났으며 X_1또는 X_2가 아닌 다른 설명변수의 빈도가 가장 높은 경우도 있다. [그림 4.1-1,2]에서 보듯이 각 설명변수의 단순 빈도수도 $p_i \neq 0.5$인 경우에 더 좋은 결과를 나타내고 있다.

[표 4.1] 설명변수가 독립인 경우 γ 의 사후분포중 빈도가 가장 높은 5개의 모형
(수자는 모형에 포함된 독립변수이고 괄호안은 모형의 비율)

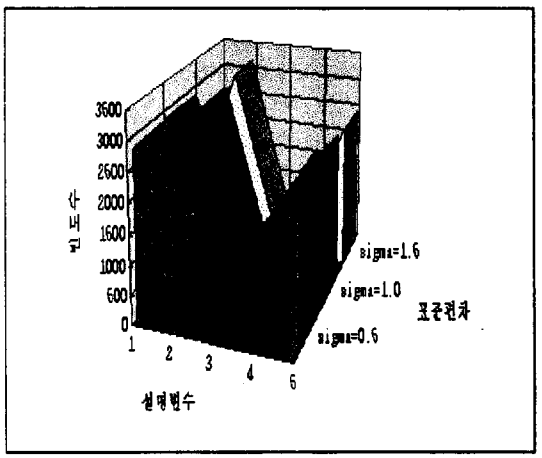
독립인 경우	$p_i = 0.5$	$0.2 < p_i < 0.8$	$0.3 < p_i < 0.7$
σ = 0.5	1 (9%)	1,2* (23.54%)	2 (13.52%)
	4 (8.3%)	2 (12.58%)	1 (10.84%)
	2 (7.9%)	1 (9.46%)	1,2* (9.7%)
	5 (6.5%)	1,2,4 (8.44%)	4 (6.3%)
	3 (6.26%)	1,2,3 (5.54%)	1,2,4 (4.5%)
σ = 1.0	3 (8.52%)	2 (13.1%)	2 (12.94%)
	1 (7.94%)	1,2* (12.08%)	1 (9.06%)
	4 (7.42%)	1,2,3 (12%)	1,2* (8.26%)
	5 (7.12%)	1 (9.48%),	3 (7.14%)
	2 (6.26%)	2,3 (7.82%)	2,3 (5.88%)
σ = 1.5	2 (11.44%)	1,2,3 (16.2%)	2 (10.44%)
	3 (8.28%)	1,2* (12.5%)	1 (8.32%)
	4 (6.72%)	2,3 (8.9%)	1,2* (7.7%)
	5 (6.66%)	2 (7.62%)	3 (6.74%)
	1 (6.26%)	1,2,3,5 (5.74%)	5 (5.56%)

[표 4.2] 설명변수가 독립이 아닌 경우 γ 의 사후분포중 빈도가 가장 높은 5개의 모형 (수자는 모형에 포함된 독립변수이고 괄호안은 모형의 비율)

독립이 아닌 경우	$p_i=0.5$	$0.2<p_i<0.8$	$0.3<p_i<0.7$
σ = 0.5	2 (8.88%)	1,2,5 (17.12%)	2 (11.18%)
	1 (8.4%)	2 (12.98%)	1 (11.14%)
	3 (7.4%)	2,5 (12.24%)	1,2* (8.92%)
	4 (6.6%)	1,2* (10.42%)	5 (7.68%)
	5 (4.96%)	1,5 (7.4%)	2,5 (7.08%)
σ = 1.0	5 (9.14%)	2 (9.8%)	2 (12.24%)
	3 (8.74%)	2,5 (9.56%)	2,5 (9.04%)
	1 (7.88%)	1,2,5 (7.22%)	5 (7.24%)
	2 (6.7%)	1,2* (7.06%),	1,2* (5.36%)
	4 (4.84%)	5 (6.72%)	2,4 (5.24%)
σ = 1.5	2 (13.78%)	2,5 (18.04%)	2 (14.3%)
	1 (8.52%)	2 (17.62%)	2,5 (9.74%)
	3 (6.92%)	5 (8.82)	5 (8.58%)
	4 (5.7%)	1,2* (5.62%)	1 (5.14%)
	5 (4.48%)	1,2,5 (5.36%)	1,2* (4.9%)



[그림 4.1-1] 설명변수가 독립이 아니고 $p=0.5$인 모의실험결과 각 설명변수의 단순빈도수



[그림 4.1-2] 설명변수가 독립이 아니고 $0.3<p<0.7$인 모의실험결과 각 설명변수의 단순빈도수

4.2 사례분석

예제 4.2 이 예제에서 사용된 실제 자료는 변수선택 문제에서 자주 인용되었던 Hald 자료(Draper and Smith, 1981) 이다. 종속변수 Y (시멘트 1그램이 화학적으로 반응하는 동안 방출하는 열)는 n = 13인 관측 값이고 p = 4 인 독립변수는 시멘트 1그램에 들어있는 화학성분들로 X_1, X_2, X_3, X_4이다. 독립변수를 표준화한 경우와 그렇지 않은 경우로 구분하여 살펴보았다. 표준화는 각 설명변수의 평균과 분산을 이용하는 일반적인 것을 사용하였다.

이 경우에도 앞의 예제에서와 비슷한 결과를 볼 수 있는데 사전 분포를 무차별 사전분포인 $p_i=0.5$ 보다 $0.2 < p_i < 0.8$ 이나 $0.3 < p_i < 0.7$로 가정한 경우가 더 최적모형에 가까운 변수를 선택할 수 있다. 여기서 최적모형은 알 수 없지만 과거의 연구(Draper and Smith, 1981)에 의하면 X_1 또는 X_2 변수가 중요한 것으로 알려져 있으므로 이를 포함하는 모형을 최적모형에 가까운 것으로 가정하였다.

표준화를 한 경우 역시 사전 분포를 무차별 사전분포인 $p_i=0.5$보다 $0.2 < p_i < 0.8$ 이나 $0.3 < p_i < 0.7$에서 더 좋은 결과를 볼 수 있다. 표준화를 하지 않은 경우를 보면, 변수를 하나도 선택하지 않은 경우의 빈도가 높은 것으로 나타나고 있는 데 이 점이 표준화를 한 경우와 다른 점이다. 결과적으로 표준화된 설명변수를 사용한 경우 더 좋은 결과를 가져왔다.

표준화를 한 경우에 $0.2<p_i<0.8$, $0.3<p_i<0.7$ 이 $p_i=0.5$인 경우보다 적합한 모형이 많은 빈도를 나타내고 있고, 표준화를 하지 않은 경우에서도 $0.2<p_i<0.8$, $0.3<p_i<0.7$ 이 $p_i=0.5$인 경우보다 회귀모형에 적합한 모형이 많은 빈도를 나타내고 있다. 각각의 변수들이 선택된 빈도의 크기는 [표 4.4]에 있다.

[표 4.3] Hald 자료에서 γ의 빈도율이 가장 높은 모형 5개 (수자는 모형에 포함된 설명변수이고 괄호안은 모형의 비율, int는 절편만 포함한 모형)

	$p_i=\frac{1}{2}$	$0.2 < p_i < 0.8$	$0.3 < p_i < 0.7$
표준화하지 않은 경우	int (21.7%)	1 (36.84%)	1 (26.68%)
	1 (14.14%)	int (16.72%)	int (21.32%)
	3 (10.28%)	1,2 (14.18%)	1,2 (9.64%)
	2 (7.62%)	2 (6.22%)	2 (7.82%)
	1,3 (6.68%)	1,3 (6.04%)	1,3 (6.38%)
표준화를 한 경우	2,3,4 (7.44%)	1,2 (20.2%)	1,2 (13.74%)
	2,3 (6.9%)	1 (18.64%)	1 (13.58%)
	1,2,4 (6.74%)	1,2,4 (8.84%)	1,2,4 (8.38%)
	3,4 (6.52%)	1,2,3 (8.34%),	1,4 (8.34%)
	1,2 (6.46%)	1,4 (8.34%)	1,2,3 (7.56%)

[표 4.4] Hald 자료에서 변수의 빈도율 크기

	$p_i=\frac{1}{2}$	$0.2 < p_i < 0.8$	$0.3 < p_i < 0.7$
표준화 하지 않은 경우	X1>X3>X2>X4	X1>X2>X3>X4	X1>X2>X3>X4
표준화를 한 경우	X2>X4>X3>X1	X1>X2>X4>X3	X1>X2>X4>X3

5. 결론

본 논문에서는 선형회귀분석에서 독립변수의 부분집합을 선택하는 한 과정으로 George와 McCulloch에 의하여 소개된 SSVS방법에서 회귀계수에 대한 사전확률을 전체모형 (full model)에서 구해진 각 계수의 t값에 따른 p-값에 의하여 결정하는 문제를 생각하였다.

결과적으로 각 설명변수가 모형에 포함되는 사전확률을 1/2로 주는 무차별 사전분포보다 p-값에 따라 선형변환된 $0.2<p_i<0.8$ 또는 $0.3<p_i<0.7$의 값을 주는 것이 다른 파라메터들(특히, c와 τ)에도 영향을 받지 않고 결과도 좋았다. 특히, 설명변수들이 독립인 경우 뿐만 아니라 회귀계수의 검정을 위한 t값이 일반적으로 왜곡된다고 알려져 있는 경우 즉, 독립변수들간에 강한 선형관계나 다중공선성이 있는 경우에도 최적모형과 필요한 변수들을 찾는 데 우리의 방법이 기존의 방법보다 효과적이었다.

George와 McCulloch는 τ와 c의 조정으로 설명변수들을 표준화한 것과 같은 효과를 얻을 수 있다고 하였는 데 Hald자료에 적용하여 본 결과 상이한 결과를 얻어 이 부분에 좀 더 자세한 연구가 필요하다고 하겠다. 또한 다른 파라메터들의 사전분포에 대한 연구도 필요하며 사전분포들이 공액(conjugate)이 아닌 경우에 깁스 표본법을 직접 적용할 수는 없지만 일반적인 마코프체인 몬테칼로(Markov Chain Monte Carlo)방법 등을 이용하여 사후분포로부터 표본수열을 얻을 수 있다면 더 넓은 부분에 사용될 수 있으며 이에 대한 연구가 진행중에 있다.

참고문헌

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, in *Second International Symposium on Information Theory.* (eds. B. N. Petrov and F. Czáki). Akademia Kiadó, Budapest, 267-81.

[2] Allen, D.M. (1974). The relationship between variable selection and data augmentation and a method for prediction, *Thechnometrics* 16, 125-127.

[3] Casella, G. and George, E.I. (1992). Explaining the Gibbs Sampler, *The American Statistician*, 46, 167-174.

[4] Draper, N., and Smith, H. (1981). *Applied Regression Analysis* (2nd ed.), New York: John Wiley.

[5] Efron, B. (1979). Bootstrap methods: Another look at the jackknife, Ann. *Statist.* 7, 1-26.

[6] Gelfand, A.E., Hills, S. E.. 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 Assocation*, 85, 972-985

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

[8] George, E.I. and McCulloch, R.E. (1993). Variable Selection Via Gibbs Sampling, *Journal of the American Statistical Association,* 88. 881-889.

[9] Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images, *IEEE Transactions on Pattern and Machine Intelligence*, 6, 721-741

[10] Hastings, W.K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications, *Biometrika*, 57, 97-109.

[11] Mallows, C.L. (1973). Some comments on C_p, *Technometrics* 15, 661-75.