

선형회귀모형에서의 효과적인 변수 선택법*

김 병 도**

《目 次》

- | | |
|-------------------|------------------|
| I. 서 론 | IV. 마케팅 데이터에의 적용 |
| II. 베이지언 변수 선택법 | V. 결 론 |
| III. 시뮬레이션에 의한 평가 | |

요 약

선형회귀모형은 어떤 현상을 예측한다든가 연구자의 가설을 검증한다든가 하는 경영학의 여러 문제에 널리 쓰이고 있다. 회귀모형을 분석하는데 있어 가장 어려운 문제 중의 하나로 지적된 분야가 변수선택의 문제이다. 즉, 종속변수를 설명하기 위한 일군의 독립변수 중에서 어떤 독립변수가 진정으로 종속변수를 설명하는데 필요한 변수인가? 그 동안 많은 연구자들은 이 독립변수 선택의 문제를 그 적용의 단순성을 이유로 단계적 회귀분석법(stepwise regression)을 사용하였다. 그러나 단계적 회귀분석법은 그 동안 여러 연구자에 의해 끊임없는 비판을 받고 있던 중, 통계학자들은 최근에 와서 변수선택의 문제를 해결하는 획기적인 방법론을 제시하기 시작하였다.

본 연구의 첫째 목적은 선형회귀모형 하에서 독립변수를 효율적으로 찾는 최근의 방법론으로 베이지언 변수선택법(Bayesian variable selection)을 소개하고 평가하는데 있다. 시뮬레이션을 통해 만들어진 여러 유형의 데이터에 단계적 회귀분석법과 베이지언 변수선택법을 적용하여 베이지언 방법론의 우수성을 보이고자 하였다. 본 연구의 둘째 목적은 모델을 선택할 때 보편적으로 사용되는 여러 척도의 문제점을 지적하고 데이터 시뮬레이션만이 진정으로 변수선택의 방법론을 평가할 수 있는 기준이 된다는 점을 강조하는데 있다. 마지막으로 실제 마케팅 데이터에 여러 변수선택법을 적용하고 각 모형을 추정하는데 이용되지 않은 샘플에

* 본 연구는 서울대학교 경영대학 경영연구소의 연구비 지원에 의해 수행되었음.

** 서울대학교 경영대학 기금 조교수

이들 모형을 적용한 결과 베이지언 변수선택법이 다른 변수선택법의 경우보다 독립변수의 수가 현저히 적은데도 불구하고 예측력에 있어서는 가장 좋은 모형의 결과를 가지게 되어 베이지언 변수선택법의 신뢰성을 한층 높이고 있다.

I. 서 론

선형회귀모형(linear regression model)을 분석하는데 있어 어떤 독립변수가 진정으로 종속변수를 설명하는데 필요한 변수인가? 이 변수선택의 문제는 실증연구를 하는 학자 그리고 예측모형을 만들고자 하는 실무자를 항상 괴롭히는 난해한 문제이다. 학자들은 변수선택의 문제를 해결하는 방법론을 크게 두 가지의 유형으로 나누고 있는데 그 첫째가 연역적(deductive)인 방법이고 둘째가 귀납적(inductive)인 접근법이다. 연역적인 방법은 실증모형을 만들기 전에 이론 연구가 선행되어야 함을 강조하고 이론을 바탕으로 독립변수를 선택하는 방법이다. 반면 귀납적인 접근법은 통계학자 또는 실증 연구자들이 선호하는 방법으로 주어진 데이터의 정보를 기초로 변수를 선택한다. 우리는 여러 상황에서 연역적인 방법이 가지는 장점을 인정하지만 임의적으로 본 논문의 주제를 귀납적인 접근법으로 국한하기로 한다.

그 동안 대다수의 연구자 그리고 실무자들은 회귀분석을 하는데 있어 독립변수 선택의 문제를 단계적 회귀분석법(stepwise regression)이라는 간편한 방법에 의존하여 해결하고자 하였다. 그러나, 단계적 회귀분석법에 의해 일군의 독립변수를 선택하는 방법은 진정한 회귀모형을 발견하는데 여러 가지 문제를 야기하였다. 단계적 회귀 분석법의 대안으로 가능한 모든 회귀분석모형을 적용하여 연구자가 사전에 정한 척도에 의해 가장 좋은 모형을 선택하는 방법론도 있지만 이는 독립변수의 수가 증대함에 따라 계산량이 기아 급수적으로 증가하여 현실성이 없다는 단점을 가지고 있다. 이에 본 연구는 독립변수를 효율적으로 찾는 방법론으로 최근에 개발되기 시작한 베이지언 변수선택법(Bayesian variable selection)을 소개하고 이를 평가하고자 한다.

본 논문이 다루고자 하는 문제를 정형화하여 표현하면 다음과 같다. 종속변수 Y 를 예측 또는 설명하기 위한 일군의 독립변수 X_1, \dots, X_p 가 주어졌다고 하자. 예를 들면, 광고액, 거시경제 지표, 판매원의 수, 제품의 가격, 등등의 독립변수의 월별 과거 데이터와 함께 제품의 월별 판매량이 주어진 시점에서 미래 판매량을 예측하기 위한 회귀분석모형을 개발하는 경우이다. 이 경우 모든 독립변수를 포함하는 선형 회귀분석모형은 다음의 행렬식 형태로 간단히 표기할 수 있다.

$$Y = X\beta + \epsilon \quad (1)$$

위의 식(1)에서 Y 는 종속변수의 관측치를 나타내는 $n \times 1$ 의 벡터이고, ϵ 은 $N(0, \sigma^2 I)$ 을 따르는 $n \times 1$ 의 오차 벡터이다. X 는 $n \times (p+1)$ 의 행렬로 첫째 행은 1로 구성된 벡터이고 나머지 p 행은 독립변수의 관측치를 나타낸다. $\beta = (\beta_0, \dots, \beta_p)'$ 는 모수 (parameter) 벡터로 β_0 은 회귀모형의 절편(intercept)을 추정하기 위한 모수이고 나머지 p 개의 β_i 는 각 독립변수의 계수를 추정하기 위한 모수이다.

주어진 회귀모형 식(1)에서 종속변수 Y 를 진실로 설명하는 일군의 독립변수를 선택하는 문제를 통계학에서는 변수선택(variable selection)의 문제라고 하는데 이를 달리 표현하면 모수 벡터 β 의 각 요소(element) β_i 가 0인가 아닌가를 결정하는 문제이다.

통계학자들은 이 변수선택의 문제를 해결하고자 과거 30년 이상 연구를 해 오고 있는데 통계학자가 권하는 한가지 방법은 가능한 모든 회귀모형을 적용하고 연구자가 사전에 정한 기준에 의해 하나의 모형을 선택하는 방법인데 이를 부분 회귀분석법(all possible subsets regression analysis)이라 부른다. 그러나 이 방법은 독립변수의 수가 많아짐에 따라 현실적으로 적용하기가 어렵다는 단점을 가지고 있다. 예를 들자면 식(1)에서 $p = 3$ 이라면 다음 8개의 서로 다른 회귀분석모형을 적용하여야 한다.

$$\emptyset, \{X_1\}, \{X_2\}, \{X_3\}, \{X_1, X_2\}, \{X_1, X_3\}, \{X_2, X_3\}, \{X_1, X_2, X_3\}$$

만약 독립변수의 수가 50개라면 약 $1.16 \times 10^{15} (= 2^{50})$ 개의 서로 다른 회귀분석모형을 적용하여야 하기 때문에 부분 회귀분석법은 매우 비효율적인 방법이라 할 수 있다. 그동안 부분 회귀분석법의 계산비용을 줄이기 위한 다수의 알고리즘이 개발되었는데 그 대표적인 알고리즘이 branch-and-bound 알고리즘이다(Furnival and Wilson 1974). 이들 알고리즘의 개발로 적용하여야 하는 회귀분석모형의 수가 현저하게 줄기는 하였지만, 아직도 많은 회귀분석모형의 문제는 부분 회귀분석법을 적용하는 것이 비현실적이다. 예로 현재 SAS, S-plus, BMDP등 대부분의 통계 분석 소프트웨어는 부분 회귀분석법을 하나의 옵션으로 가지고 있지만 독립변수의 수를 최고 30개 정도로 제한하거나 그 수가 30개 이상인 경우 계산시간이 현저히 증가하고 있다.

부분 회귀분석법이 가지는 또 하나의 단점은 가장 좋은 모형을 선택하기 위해 어떤 척도를

사용하여야 하는가에 대한 문제이다. 현재 보편적으로 많이 쓰이고 있는 Adjusted R^2 나 AIC (Akaike Information Criteria) 등의 척도는 진정한 모형을 찾는데 한계가 있음을 본 논문은 보이고 있다. 제 3장에서 시뮬레이션을 통해 보다 자세히 보이겠지만, 진정한 모형을 선택하는 척도가 불완전하다면 부분 회귀분석법으로 선택된 모형은 진정한 모형이 아닐 가능성이 높을 것이라는 것은 당연한 논리이다.

만약 독립변수의 수가 30개 이상의 경우는 어떠한 방법으로 변수를 선택하여야 하는가? 이 경우 대다수의 연구자들은 적용의 단순성을 이유로 단계적 회귀분석법을 이용한다. 단계적 회귀분석법은 휴리스틱한 방법에 의존하여 부분 회귀분석법이 가지는 계산상의 부담을 해결한 것은 사실이지만, 많은 연구자들에 의해 끊임없는 비판의 대상이 되어 오고 있다 (Hocking 1976; Draper and Smith 1982; Miller 1984). 단계적 회귀분석법이 가지는 가장 큰 단점은 각 단계에서 하나의 변수를 채택하고 기각하는데 필요한 유의수준 자체가 주어진 데이터의 특성에 따라 달라져야 한다는데 있다. 또한, 독립변수 상호간에 상관계수가 높은 경우 단계적 회귀분석법은 좋은 성과를 내지 못함이 지적되었다 (Berk 1978).

II. 베이지언 변수 선택법 (Bayesian Variable Selection)

최근 통계학계에서는 단계적 회귀분석법이 가지는 이론적 단점을 보완함과 동시에 부분 회귀분석법이 가지는 계산상의 문제를 해결하는 변수 선택의 새로운 방법론을 제시되기 시작하였다 (George and McCulloch 1993; Smith and Kohn 1996). 이 새로운 방법론은 변수선택의 문제를 베이지언 혼합모형(Bayesian mixture model)에서 부분집합을 찾는 문제로 접근하는데 이 경우 이후화률(posterior probability)이 큰 부분집합이 우리가 찾는 진정한 모형이 된다는 논리이다. 또한, 각 모형의 이후화률의 분포를 직접적으로 계산하는데 따르는 어려움을 극복하기 위해 갑스샘플러(Gibbs sampler) 기법을 사용한다는 점이 이 방법론의 핵심이 되는 부분이다. 갑스샘플러는 최근 컴퓨터의 계산능력의 급속한 향상과 더불어 등장한 대표적인 통계적 방법론으로 그 동안 베이지언 통계학에서 해결하지 못했던 여러 가지 문제에 해법을 제시하면서 주목할 받기 시작한 방법론이다.

식 (1)의 회귀분석모형으로 되돌아가서 종속변수 Y 를 진정으로 설명하는 일군의 독립변수를 선택하는 문제는 모두 벡터 β 에서 0이 아닌 일군의 β_i 를 발견하는 문제와 동일하다고 할 수 있다. 표기의 편의상 $\gamma_i = 0$ if $\beta_i = 0$ 이고 $\gamma_i = 1$ if $\beta_i \neq 0$ 인 $(p+1) \times 1$ 의 벡터 $\gamma = (\gamma_0, \dots, \gamma_p)'$ 를 정의하기로 하자. 또한 주어진 벡터 γ 에 대하여 β_γ 는 0이 아닌 β_i 들

로 구성된 벡터라고 정의하고, X_γ 는 $\gamma_i = 1$ 인 항에 상응하는 X 의 행으로 구성된 행렬이라고 하자. 예를 들어 γ 벡터가 $(1, 0, \dots, 0)'$ 으로 주어졌다면 $\beta_\gamma = \beta_0$ 이고 $X_\gamma = X_0$ 이 됨을 의미한다. 즉, 절편만을 가지는 회귀모형을 의미한다.

변수선택의 문제를 베이지언 통계학의 틀에서 접근하기 위해 우리는 먼저 각 모수의 이전 확률(prior probability) 분포에 대한 가정을 하여야 한다. 우리가 추정하고자 하는 모수는 β_γ 와 σ^2 그리고 γ 인데 이들의 결합확률분포에 대한 이전 확률분포 $\pi(\beta_\gamma, \sigma^2, \gamma)$ 는 베이즈의 정리에 따라 $\pi(\beta_\gamma | \sigma^2, \gamma) \pi(\sigma^2 | \gamma) \pi(\gamma)$ 의 세 개의 조건부확률의 곱과 동일하다. 그러므로 이들 각 조건부확률에 대한 가정을 서술하기로 한다. 첫째, β_γ 의 이전확률분포는

$$\pi(\beta_\gamma | \sigma^2, \gamma) \sim N(0, c\sigma^2(X_\gamma'X_\gamma)^{-1}) \quad (2)$$

위의 식(2)에서 c 의 값은 연구자가 임의로 설정하는 상수값으로, Smith와 Kohn (1996)은 광범위한 시뮬레이션을 통해 10보다 크고 1000보다 적은 수 중에 하나를 선택하면 무방하다고 주장하고 있다. c 의 값의 크기는 결국 β_γ 를 추정하는데 있어 사전정보의 양과 데이터가 제공하는 정보의 양의 기여도에 의해 결정되는 것으로, c 의 값을 10보다 크게 설정하였다는 의미는 β_γ 에 대한 사전정보가 거의 없다고 가정함을 의미한다. 우리도 10보다 큰 여러 c 값을 적용하여 보았으나 결과에 있어 거의 차이를 보이지 않았기 때문에 $c = 100$ 인 경우의 결과만을 보고하기로 한다.

β_γ 는 어떤 형태의 이전확률분포를 가정하는 것이 타당한가에 대하여 학자들간에 이견이 존재하는데 (George and McCulloch 1996), 식(2)의 형태는 Smith와 Kohn (1996)가 채택한 가정으로 Raftery, Madigan과 Hoeting (1993)도 유사한 가정을 하고 있다. 우리가 식(2)의 가정을 채택한 이유는 상대적으로 이후확률분포(posterior distribution)의 도출이 단순하고 변수를 추정하는 속도가 빠르다는데 있다.

둘째, σ^2 의 이전확률분포는 제프리(Jeffrey)의 주장에 따라 $\pi(\sigma^2 | \gamma) \propto 1/\sigma^2$ 을 따른다고 가정한다 (Zellner 1971). 이와 같은 가정은 σ^2 의 이전확률분포로 베이지언 통계분석법에서 통상적으로 쓰이는 가정으로 이는 $\log(\sigma^2)$ 가 균일분포(uniform distribution)라는 가정과 동일한 가정이다. 이 가정은 σ^2 를 추정하는데 있어 사전정보의 영향을 최소화하고 데이터의 기여도를 극대화한다는 점에 있어서 설득력이 있는 가정이라 할 수 있다.

셋째, γ 벡터의 이전확률 분포에 대한 가정인데 우리는 각 γ_i 는 서로 독립적이고 각 독립 변수가 진정으로 종속변수 Y 를 설명하는가에 대한 어떤 사전 정보도 없다고 가정하여 $p(\gamma_i = 1) = p(\gamma_i = 0) = \frac{1}{2}$ 이라고 가정한다. 이 가정하에서 γ 는 벡터의 차원(dimension)이 $p+1$ 이므로 벡터 γ 의 결합확률분포 $p(\gamma)$ 는 $2^{-(p+1)}$ 이 된다. 물론 연구자가 과거의 경험이나 주관에 의해 어떤 독립변수가 종속변수를 설명하는 정도를 알고 있다면 이를 γ 벡터의 이전확률분포에 쉽게 반영할 수 있다.

우리는 베이즈의 정리에 의하여 위에서 가정한 각 모수의 이전확률분포와 데이터로부터의 정보를 결합하여 모수의 이후확률분포를 다음의 식으로부터 도출할 수 있다.

$$\begin{aligned} p(\beta_\gamma, \sigma^2, \gamma | Y) &\propto p(Y | \beta_\gamma, \sigma^2, \gamma) \pi(\beta_\gamma, \sigma^2, \gamma) \\ &= p(Y | \beta_\gamma, \sigma^2, \gamma) \pi(\beta_\gamma | \sigma^2, \gamma) \pi(\sigma^2 | \gamma) \pi(\gamma) \end{aligned} \quad (3)$$

우리는 식(3)으로부터 변수선택 문제의 핵심인 γ 의 이후확률분포 $p(\gamma | Y)$ 을 다음과 같이 도출할 수 있다.

$$\begin{aligned} p(\gamma | Y) &\propto p(Y | \gamma) \pi(\gamma) \\ &= \left[\int_{\sigma^2} \left[\int_{\beta_\gamma} p(Y | \beta_\gamma, \sigma^2, \gamma) \pi(\beta_\gamma | \sigma^2, \gamma) d\beta_\gamma \right] \pi(\sigma^2 | \gamma) d\sigma^2 \right] \pi(\gamma) \\ &\propto (1 + c)^{-q_\gamma/2} S(\gamma)^{-n/2} \pi(\gamma) \\ &= (1 + c)^{-q_\gamma/2} S(\gamma)^{-n/2} 2^{-(p+1)} \end{aligned} \quad (4)$$

where $S(\gamma) = Y'Y - \frac{c}{1+c} Y'X_\gamma (X_\gamma' X_\gamma)^{-1} X_\gamma' Y$ and $q_\gamma = \sum_{i=0}^p \gamma_i$

식(4)로부터 우리는 γ 의 이후확률분포를 분석적으로 도출할 수 있고 그 결과 모수 γ 에 대한 추정이 가능하다. 즉, γ 의 최빈값 γ_{mode} 로부터 어떤 독립변수를 선택하여야 하는가에 대한 결정을 하는 것이 하나의 추정 방법이 될 수 있다. 또는 모형 자체에 대한 불확실성을 고려할 수도 있는데 이의 예로 γ 의 차원이 2이라 하자. 우리는 식(4)에서 $p(\gamma_0 = 0, \gamma_1 = 0), p(\gamma_0 = 1, \gamma_1 = 0), p(\gamma_0 = 0, \gamma_1 = 1), p(\gamma_0 = 1, \gamma_1 = 1)$ 을 계산할 수 있고 이는 우리가 찾는 모형이 $Y = 0$ 일 확률이 $p(\gamma_0 = 0, \gamma_1 = 0), Y = \beta_0$ 일 확률이 $p(\gamma_0 = 1, \gamma_1 = 0), Y = \beta_1 X_1$ 일 확률이 $p(\gamma_0 = 0, \gamma_1 = 1)$, 그리고 $Y = \beta_0 + \beta_1 X_1$ 일 확률이 $p(\gamma_0 = 1, \gamma_1 = 1)$ 이

라는 추론을 할 수 있다. 그러나 이와 같은 γ 의 이후 확률분포를 분석적으로 도출하는 일은 독립 변수의 수가 큰 경우 현실적으로 그 계산이 불가능하다. 예를 들어 $p = 100$ 인 경우 우리는 2^{100} 개의 케이스를 계산하여야 한다.

이 문제를 해결하기 위해 우리는 γ 의 이후 확률분포를 분석적으로 도출하는 대신에 킁스샘플러라 불리는 간접적인 방법을 이용한다 (Tanner and Wong 1987; Gelfand and Smith 1990). 즉, 주어진 γ 의 이후 확률분포 함수로부터 무작위로 많은 수 그러나 현실적으로 가능한 수의 $\gamma^{[i]}$ 를 추출하여 γ 의 이후 확률분포 함수를 추정하여 보는 것이다. 위의 2 차원의 γ 의 예로 돌아가서 주어진 식(4)로부터 만약 10,000회 γ 를 추출하였더니 $\gamma = (\gamma_0 = 0, \gamma_1 = 0)$ 인 경우가 500회, $\gamma = (\gamma_0 = 0, \gamma_1 = 1)$ 인 경우가 300회, $\gamma = (\gamma_0 = 1, \gamma_1 = 0)$ 인 경우가 100회, $\gamma = (\gamma_0 = 1, \gamma_1 = 1)$ 인 경우가 100회 실현되었다면 우리는 $p(\gamma_0 = 0, \gamma_1 = 0) = \frac{1}{2}, p(\gamma_0 = 1, \gamma_1 = 0) = 3/10, p(\gamma_0 = 0, \gamma_1 = 1) = p(\gamma_0 = 1, \gamma_1 = 1) = 1/10$ 이라고 추정하는 것이다. 이와 같은 킁스샘플러의 방법론을 이용하면 γ 의 차원이 클 경우 연구자 자신이 무작위 추출의 수를 정할 수 있기 때문에 매우 유용하다.

식(4)에서 주어진 분포로부터 γ 를 추출하는 킁스샘플러를 보다 상세히 설명하면 다음과 같다 (Smith and Kohn 1996; George and McCulloch 1996).

1단계: γ 의 초기값 $\gamma^{[0]} = (\gamma_1^{[0]}, \dots, \gamma_p^{[0]})$ 을 선택하여야 하는데, 각 $\gamma_i^{[0]}$ 의 값을 $p(\gamma_i = 1) = p(\gamma_i = 0) = \frac{1}{2}$ 인 이항분포에서 무작위로 추출된 값으로 정하면 무방하다.

2 단계: 각 $\gamma_i (i = 0, 1, \dots, p)$ 의 값을 조건부 확률분포 $p(\gamma_i | Y, \gamma_{j \neq i})$ 로부터 순차적으로 무작위 추출한다. 식(4)로부터 조건부 확률 $p(\gamma_i | Y, \gamma_{j \neq i})$ 은 다음과 같이 도출할 수 있다.

$$p(\gamma_i | Y, \gamma_{j \neq i}) \propto p(Y | \gamma) \pi(\gamma_i) \propto (1 + c)^{-q_i/2} S(\gamma)^{-n/2} 2^{-1} \quad (5)$$

III. 시뮬레이션에 의한 평가

지금까지 설명한 베이지언 변수선택 방법론이 단계적 회귀분석법과 같은 기존의 방법론에 비해 어느 정도 우수한 방법론인가를 평가하기 위해 우리는 시뮬레이션에 의해 3개의 인위적 데이터를 만들기로 한다. 첫번째 데이터는 서로 독립적인 50개의 독립변수를 표준정규분포

$N(0, 1)$ 에서 100개씩 추출한다. 즉, X_1, \dots, X_{50} i.i.d. $N_{100}(0, 1)$. 그리고 ϵ 을 독립적으로 표준정규분포 $N(0, 1)$ 에서 100개 추출하고 우리는 $Y = \epsilon$ 이라 가정한다. 즉, 종속변수 Y 를 진정으로 설명하는 회귀식은 위의 50개의 독립변수 중 어느 것도 포함하지 않아야 한다는 모형이다. 이 상황에서 진정한 회귀모형을 발견하기 위해 연구자들이 보편적으로 쓰는 방법인 단계적 회귀모형을 적용하여 보기로 하자. 변수 도입유의수준과 탈락유의수준 모두를 0.05로 정한 단계적 회귀모형을 위의 데이터에 적용하였을 때 선택된 모형의 결과를 요약하면 다음과 같다.

〈표 1〉 시뮬레이션 1: 단계적 회귀분석법

| 모 수 | 추정치(표준편차) | p value |
|--------------|-------------|---------|
| β_0 | 0.02(0.10) | 0.828 |
| β_2 | -0.20(0.10) | 0.045 |
| β_{27} | -0.31(0.11) | 0.006 |
| β_{36} | -0.22(0.10) | 0.030 |
| β_{45} | 0.35(0.09) | 0.000 |

진정한 모형은 어떤 독립변수도 선택하지 말아야 하는데도 불구하고 단계적 회귀모형은 4개 ($X_2, X_{27}, X_{36}, X_{45}$)의 독립변수를 선택하였다. 그리고 그들 모두의 유의수준도 $\alpha = 0.05$ 에서 유의하였고 위 회귀식의 적합도 (R^2) 역시 0.18로 유의하였다. 즉, 단계적 회귀분석법을 적용하였을 때 실제는 아무 설명력이 없어야 하는 다수의 독립변수가 선택되었다.

동일한 데이터에 베이지언 변수선택법을 적용하기로 하자. 베이지언의 경우는 변수선택의 문제를 위에서 설명한 바와 같이 킁스샘플링에 의해 γ 의 이후확률분포에서 다수의 벡터 γ 를 무작위로 추출하는데, 우리는 벡터 γ 를 10,000회 추출하기로 한다. 이를 요약하여 10,000회에서 가장 빈번하게 발생한 10개의 벡터 γ 를 보면 다음과 같다.

〈표 2〉 시뮬레이션 1: 베이지언 변수선택법

| 선택된 변수 | 빈도 |
|------------------|-----|
| 아무 변수도 선택되지 않음 | 354 |
| X_{45} | 196 |
| X_{27}, X_{45} | 85 |
| X_{21} | 80 |
| X_{43} | 64 |
| X_{36}, X_{45} | 52 |
| X_{50} | 47 |
| X_{27} | 38 |
| X_{21}, X_{45} | 37 |
| X_{36} | 36 |

〈표 2〉에서 볼 수 있는 가장 흥미로운 사실은 어떤 변수도 선택되지 않은 모델이 가장 빈번하게 발생하였다는 점이다. 즉, 우리가 벡터 γ 의 이후확률분포로부터 하나의 모형을 선택한다면 분포의 최빈값(Mode)으로 선택할 수 있는데 벡터 γ 의 최빈값은 “어떤 변수도 선택하지 않는 모형”이 된다는 것이다. 단계적 변수선택법은 동일한 데이터에서 4개의 변수를 선택하였다는 점을 고려할 때, 위의 결과는 베이지언 변수선택법이 진정한 모형을 찾는데 얼마나 유용한 방법론인가를 잘 보여 준 예라고 할 수 있다. 둘째로, 베이지언 변수 선택법은 변수 선택에 있어서의 불확실성을 확률분포의 개념으로 표현하고 있다. 즉, 주어진 데이터로부터 우리가 추론할 수 있는 진정한 모형을 찾는 데는 불확실성이 따른다는 사실인데, 위의 표에서 보면 다른 모형들의 케이스도 빈번하게 발생하였다는 점을 주목할 필요가 있다. 그러나 그와 같은 차선의 모형들도 두개 이하의 독립변수를 포함하고 있어 우리가 표본오차에 의해 차선의 모형을 선택하였다 할 지라도 최소한 단계식 회귀분석법의 경우보다는 나은 모형을 선택한다는 점이 고무적이다.

두 번째 시뮬레이션은 첫 번째 시뮬레이션과 매우 유사하다. 서로 독립적인 50개의 독립변수 X_1, \dots, X_{50} 을 표준정규분포 $N(0, 1)$ 에서 100개씩 추출한다. 그리고 ε 을 독립적으로 표준정규분포 $N(0, 1)$ 에서 100개 추출하고 Y 는 $Y = X_1 + 2X_2 + \varepsilon$ 의 값을 가진다고 정의 한다. 이렇게 형성된 데이터를 이용하여 먼저 종속변수 Y 를 진정으로 설명하는 일군의 독립변수를 찾기 위해 먼저 단계적 회귀모형을 적용한다. 첫 번째 시뮬레이션과 마찬가지로 변수 도입유의수준과 탈락유의수준 모두를 0.05로 정한 단계적 회귀모형을 적용하였을 때 선택된 모형의 결과를 요약하면 다음과 같다.

〈표 3〉 시뮬레이션 2: 단계적 회귀분석법

| 모 수 | 추정치(표준편차) | p value |
|--------------|-------------|---------|
| β_0 | -0.06(0.09) | 0.495 |
| β_1 | 0.86(0.08) | 0.000 |
| β_2 | 1.95(0.10) | 0.000 |
| β_{11} | 0.21(0.09) | 0.018 |
| β_{14} | -0.19(0.08) | 0.018 |
| β_{28} | 0.23(0.09) | 0.018 |
| β_{38} | 0.22(0.09) | 0.015 |
| β_{42} | -0.21(0.09) | 0.028 |
| β_{50} | 0.20(0.09) | 0.029 |

위 회귀식의 적합도 (R^2)는 0.85로 유의하였는데 진실로 종속변수를 설명하는 2개의 독립

변수를 찾아내었다. 그러나, 단계적 회귀분석법은 6개 ($X_{11}, X_{14}, X_{28}, X_{38}, X_{42}, X_{50}$)의 허구적인 독립변수를 선택하고 이를 모두가 유의수준 $\alpha = 0.05$ 에서 유의하였다.

이제 동일한 데이터에 베이지언 변수선택법을 적용하여 그 결과를 단계적 회귀분석법과 비교하기로 하자. γ 의 이후확률분포로부터 벡터 γ 를 10,000회 무작위 추출하여 가장 빈번하게 발생하는 10개의 벡터 γ 를 정리하면 다음과 같다.

〈표 4〉 시뮬레이션 2: 베이지언 변수선택법

| 선택된 변수 | 빈도 |
|------------------------------------|-----|
| X_1, X_2 | 117 |
| X_1, X_2, X_{11}, X_{50} | 81 |
| X_1, X_2, X_{50} | 78 |
| X_1, X_2, X_{38} | 73 |
| X_1, X_2, X_{38}, X_{50} | 72 |
| X_1, X_2, X_{11} | 44 |
| X_1, X_2, X_{28}, X_{38} | 34 |
| X_1, X_2, X_{28}, X_{50} | 31 |
| $X_1, X_2, X_{14}, X_{28}, X_{38}$ | 31 |
| X_1, X_2, X_{14}, X_{50} | 30 |

첫 번째 시뮬레이션의 경우와 마찬가지로 벡터 γ 의 최빈값은 X_1 과 X_2 만을 선택한 모형으로 베이지언 변수선택법으로 진정한 모형을 발견할 수 있었다. 또한, 단계적 변수 선택법은 6개의 허구의 변수를 선택한 반면, 베이지언 변수 분석법의 경우는 차선의 모형들도 1~3개의 허구 변수를 선택하였다는 점을 주목할 필요가 있다. 즉, 베이지언 변수 분석법의 경우가 진실한 모형과 가까운 모형을 발견할 확률이 높다는 사실이다.

베이지언 변수선택법의 우수성을 보이기 위하여 우리가 수행한 마지막 시뮬레이션은 독립 변수 간에 상관계수가 높은 경우이다. 과거의 연구에 의하면 독립변수 간에 상관계수가 높은 경우 단계식 변수 선택법이 변수선택을 하는데 문제가 있다는 점이 지적되었다 (Berk 1978). 우선 서로 독립적인 51개의 독립변수 Z_1, \dots, Z_{51} 을 표준정규분포 $N(0,1)$ 에서 100 개씩 추출한다. 그리고 $X_i = Z_i + 2Z_{51} (i = 1, \dots, 50)$ 이라고 정의하면 X_i 간의 상호 상관계수는 약 0.8정도 되는 50개의 독립 변수의 벡터를 얻을 수 있다. 마지막으로 ϵ 을 독립적으로 표준정규분포 $N(0,1)$ 에서 100개 추출하고 Y 는 $Y = X_1 + 2X_2 + \epsilon$ 이라고 정의한다. 이렇게 형성된 데이터를 이용하여 먼저 종속변수 Y 를 진정으로 설명하는 일군의 독립변수를 찾기 위해 먼저 단계적 회귀모형을 적용한다. 위의 두 시뮬레이션과 마찬가지로 변수 도입유의수준과 탈락유의수준 모두를 0.05로 정한 단계적 회귀모형을 적용하였을 때 선택된 모형의 결과를 요약하면 다음과 같다.

〈표 5〉 시뮬레이션 3: 단계적 회귀분석법

| 모 수 | 추정치(표준편차) | p value |
|--------------|-------------|---------|
| β_0 | 0.23(0.08) | 0.005 |
| β_1 | 0.87(0.08) | 0.000 |
| β_2 | 2.13(0.08) | 0.000 |
| β_7 | 0.21(0.06) | 0.001 |
| β_{20} | 0.19(0.08) | 0.027 |
| β_{23} | -0.24(0.08) | 0.004 |
| β_{42} | -0.15(0.06) | 0.024 |

〈표 5〉의 회귀모형의 적합도는 0.99로 매우 높았고 종속변수를 진실로 설명하는 2개의 독립 변수를 선택하였다. 그러나 단계적 회귀모형은 유의수준 $\alpha = 0.05$ 에서 유의한 4개 ($X_7, X_{20}, X_{23}, X_{42}$)의 허구적인 독립변수를 선택하였다. 여기서 하나 주의할 점은 진정한 모델은 절편 (intercept)이 없는 반면 위의 모형은 절편이 통계적으로 유의하다는 점이다. 즉, 우리는 절편의 선택도 변수선택의 문제로 볼 수 있기 때문에 단계적 회귀모형은 5개의 허구적인 독립 변수를 선택하였다고 할 수 있다. 그러나 과거의 연구에서 지적한 바와는 다르게 독립변수가 서로 상관되어 있는 경우 특별히 단계적 회귀분석법이 문제가 발생하는 것은 아니었다. 우리는 어떤 상황에서 단계적 회귀분석법이 진정한 모형을 찾는데 더 어려움을 가지는가에 대한 보다 심층적인 연구의 필요성을 강조하다.

세 번째 데이터에 베이지언 변수선택법을 적용하기로 하자. 위의 두 경우와 마찬가지로 벡터 γ 를 10,000회 추출하여 가장 빈번하게 발생하는 10개의 벡터 γ 를 정리하면 다음과 같다.

〈표 6〉 시뮬레이션 3: 베이지언 변수선택법

| 선택된 변수 | 빈도 |
|--------------------------------------|-----|
| X_1, X_2 | 159 |
| $X_0, X_1, X_2, X_7, X_{23}$ | 114 |
| X_1, X_2, X_7, X_{42} | 97 |
| X_1, X_2, X_7, X_{23} | 63 |
| $X_1, X_2, X_5, X_7, X_{42}$ | 58 |
| X_0, X_1, X_2 | 58 |
| $X_0, X_1, X_2, X_7, X_{42}$ | 58 |
| X_1, X_2, X_7 | 49 |
| X_0, X_1, X_2, X_7 | 46 |
| $X_0, X_1, X_2, X_7, X_{23}, X_{42}$ | 42 |

이전의 두 시뮬레이션의 경우와 마찬가지로 벡터 γ 의 최빈값은 X_1 과 X_2 만을 선택한 모형으로 베이지언 변수선택법은 정확히 진실한 모형을 찾을 수 있었다. 즉, 베이지언 변수선택

법은 독립변수 간의 상관관계에 유무에 관계없이 진정한 모형을 찾아내었다. 또한, 단계적 변수 선택법은 4개의 허구의 변수를 (절편을 포함하면 5개) 선택한 반면, 베이지언 변수 분석법의 경우는 차선의 모형들도 1~4개의 허구 변수를 선택하였다. 끝으로 우리는 벡터 γ 의 이후확률분포를 알기 위해 10,000회의 벡터 γ 를 추출한 후 그 중에서 가장 빈번하게 발생하는 10개의 경우만을 <표 6>에 요약하였다. 즉, 10,000회 시행 중에서 744회의 경우가 <표 6>에서 요약된 벡터 γ 가 발생하였고 90% 이상의 경우는 다른 사건이 발생하였다는 것이다. 비록 각각의 사건이 발생할 확률은 매우 적지만 이들을 합한 확률은 매우 크다는 것이다. 이들 10,000회의 시행된 벡터 γ 를 보다 면밀히 살펴보면 X_1 과 X_2 는 10,000회 모든 케이스에서 독립변수로 포함되었다. 반면 $X_0, X_3, X_7, X_{23}, X_{42}$ 는 10퍼센트 이상의 케이스에서 독립변수로 포함되었으며 그 나머지 변수들은 독립변수로 포함될 확률이 10퍼센트 미만이었다. 우리가 여기서 강조하고자 하는 점은 주어진 데이터로부터 진정한 모형을 발견하는 문제에는 항상 불확실성이 존재하는 하나의 추정 문제라는 점이다. 이런 불확실성 또는 오류를 줄이는 위해 데이터의 수를 늘리는 것이 하나의 방법이 될 수 있다. 주어진 데이터가 가진 정보의 활용을 극대화하는 방법론을 사용하는 것이 또 다른 방편이 될 수 있는데 베이지언 변수선택법은 기존의 단계적 회귀분석법보다 이 점에 있어 진일보된 방법론이라는 것이다.

IV. 마케팅 데이터에의 적용

베이지언 변수선택 방법론을 보다 현실적인 상황에서 평가하기 위하여 IRI (Information Resources, Inc.)가 제공한 데이터에 적용하였다. 미국의 대표적인 시장조사 업체인 IRI는 여러 종류의 스캐너 데이터를 수집하여 소매점과 제조업체에 제공하고 있는데 우리는 그 중에서 별로 마케팅 연구자가 사용하지 않은 소비자의 구매액과 관련된 스캐너 데이터를 분석하고자 한다. 이 데이터는 1984년 1월부터 약 2년 동안 1,398명의 소비자가 슈퍼에서 얼마를 지출하였고 그들이 어떤 인구통계학적 특성을 가지고 있는가를 기록하고 있다. (데이터에 대한 보다 자세한 설명을 위해서는 Kim과 Park (1997)의 논문을 보라.)

잠재시장에서 우수고객을 발견하는 일은 슈퍼의 마케팅 관리자에게는 대단히 중요한 업무의 하나이다. 예를 들어 슈퍼에서의 평균지출액이 많은 고객을 발굴하여 이를 우수회원을 위한 특별사은행사를 할 수 있다. 즉, 어떤 인구통계학적 특성을 갖는 고객이 우수고객인가를 예측하는 일은 시장세분화의 근간이 된다. 우리는 여기서 소비자의 슈퍼에서의 지출액에 대한 이론적 모형을 세우기보다는 주어진 인구통계적 변수를 이용하여 지출액을 합리적으로 설

명 또는 예측할 수 있는 선형회귀모형을 추정하고자 한다. IRI가 1,398명의 소비자에 대하여 수집한 매주 평균 지출액과 18개의 인구통계적 변수는 다음과 같다.

〈표 7〉 IRI 데이터 변수의 개괄적 설명

| 변 수 | 설 명 | 평균값* |
|-----------------------|---------------------------|----------|
| 주간 지출액(Y) | 평균적으로 일주일에 슈퍼에서 지출하는 금액 | \$47.57 |
| 가구구성원 수(X_1) | 가구 구성원의 수 | 2.83명 |
| 소득(X_2) | 연평균 소득 | \$26,188 |
| 인종(X_3) | 백인이면 1이고 유색 인종이면 0 | 0.88 |
| 주거지역(X_4) | 도시에 거주하면 1 시외에 거주하면 0 | 0.63 |
| 주택의 종류(X_5) | 단독주택에 거주하면 1 그 외는 0 | 0.84 |
| 주택의 소유(X_6) | 주택을 소유하고 있으면 1 아니면 0 | 0.83 |
| 개의 소유(X_7) | 개를 기르고 있으면 1 아니면 0 | 0.45 |
| 고양이의 소유(X_8) | 고양이를 기르고 있으면 1 아니면 0 | 0.26 |
| 어린이의 유무(X_9) | 12세 미만의 어린이가 있으면 1 아니면 0 | 0.35 |
| 전업직장의 유무(X_{10}) | 가장이 전업직업을 가지고 있으면 1 아니면 0 | 0.50 |
| 대학교육(X_{11}) | 가장이 대학졸업 이상의 학력이면 1 아니면 0 | 0.20 |
| 토스터의 유무(X_{12}) | 토스터를 소유하고 있으면 1 아니면 0 | 0.88 |
| 커피메이커의 유무(X_{13}) | 커피메이커를 소유하고 있으면 1 아니면 0 | 0.69 |
| 식기세척기의 유무(X_{14}) | 식기세척기를 소유하고 있으면 1 아니면 0 | 0.54 |
| 전자랜지의 유무(X_{15}) | 전자랜지를 소유하고 있으면 1 아니면 0 | 0.65 |
| 개인용컴퓨터 유무(X_{16}) | 개인용컴퓨터를 소유하고 있으면 1 아니면 0 | 0.16 |
| 세탁기의 유무(X_{17}) | 세탁기를 소유하고 있으면 1 아니면 0 | 0.79 |
| VTR의 유무(X_{18}) | VTR을 소유하고 있으면 1 아니면 0 | 0.39 |

* 위의 평균치는 각 변수에 대하여 총 1,398명의 산술평균값을 의미함

우리는 1,398명의 소비자를 무작위로 1,000명과 398명의 두 집단으로 나누고, 1,000명의 소비자 샘플에 변수선택법의 적용과 함께 회귀분석모형을 적용하여 모수를 추정한다. 그리고, 나머지 398명의 소비자에게 추정된 모형을 적용하여 모형의 예측력을 테스트하기로 한다. 즉, 우리는 모형의 적합성을 측정하는데 있어 샘플내 테스트와 샘플외 테스트를 병행한다는 것이다.

우리는 먼저 1,000명의 소비자 데이터에 단계적 회귀모형을 적용한다. 18개의 독립변수로 시작하여 변수 도입유의수준과 탈락유의수준 모두를 0.05로 정한 단계적 회귀모형을 적용하여 선택된 모형을 요약하면 다음과 같다.

〈표 8〉 IRI 데이터: 단계적 회귀분석법

| 모 수 | 추정치(표준편차) | p value |
|--------------|--------------|---------|
| β_0 | -16.90(6.51) | 0.010 |
| β_1 | 5.22(0.50) | 0.000 |
| β_2 | 3.36(0.69) | 0.000 |
| β_4 | 4.09(1.29) | 0.002 |
| β_5 | 4.43(1.78) | 0.013 |
| β_7 | 6.17(1.29) | 0.000 |
| β_8 | 4.61(1.46) | 0.002 |
| β_{11} | 3.57(1.64) | 0.029 |
| β_{13} | 3.67(1.39) | 0.008 |
| β_{15} | 3.40(1.44) | 0.019 |
| β_{18} | 2.99(1.38) | 0.030 |

〈표 8〉의 회귀모형의 적합도는 0.29이고 절편을 포함하여 추정된 11개의 모수 모두가 유의 수준 $\alpha = 0.05$ 에서 통계적으로 유의하였다.

단계적 변수선택법과의 비교를 위하여 동일한 데이터에 베이지언 변수선택법을 적용하기로 한다. 시뮬레이션의 경우와 마찬가지로 γ 의 이후확률분포로부터 벡터 γ 를 10,000회 무작위 추출하여 가장 빈번하게 발생하는 10개의 벡터 γ 를 정리하면 다음과 같다.

〈표 9〉 IRI 데이터: 베이지언 변수선택법

| 선택된 변수 | 빈도 |
|---|-----|
| $X_0, X_1, X_2, X_4, X_5, X_7, X_8, X_{15}$ | 206 |
| $X_0, X_1, X_2, X_7, X_8, X_{11}, X_{15}$ | 176 |
| $X_0, X_1, X_2, X_4, X_5, X_7, X_8, X_{13}, X_{15}$ | 165 |
| $X_0, X_1, X_2, X_4, X_7, X_8, X_{15}$ | 163 |
| $X_1, X_2, X_7, X_8, X_{11}, X_{13}, X_{15}$ | 138 |
| $X_1, X_2, X_7, X_8, X_{11}, X_{15}, X_{18}$ | 127 |
| $X_0, X_1, X_2, X_4, X_7, X_8, X_{15}, X_{18}$ | 117 |
| $X_0, X_1, X_2, X_3, X_4, X_5, X_7, X_8, X_{15}$ | 111 |
| $X_1, X_2, X_4, X_7, X_8, X_{11}, X_{15}$ | 107 |
| $X_0, X_1, X_2, X_4, X_7, X_8, X_{14}, X_{16}$ | 102 |

γ_{mod} 즉 가장 빈번히 발생한 벡터 γ 를 선택하면 이는 절편을 포함하여 8개의 변수를 가진 모형이다. 반면 단계적 회귀분석법은 절편을 포함하여 11개의 변수를 선택하였다. 어떤 모형이 진정한 회귀모형인가? 시뮬레이션의 경우는 사전에 어떤 변수가 진정으로 설명력이 있는가를 알고 있기 때문에 방법론의 비교가 용이하지만, 실제 데이터의 적용에서는 방법론의 비교가 그렇게 단순하지 않다. 한가지 간접적인 방법은 여러 가지 척도에 의해 각 모델을 비교

하는 것인데 그 대표적인 척도의 예가 R^2 , Adjusted R^2 , AIC (Akaike Information Criteria) 등이다. 그러므로 우리는 이 세 가지 척도에 의해 두 모형을 비교하고자 한다. 또한 독립변수의 수가 18개이므로 부분회귀분석 즉 모든 가능한 회귀모형을 시행하여 그 중에서 각 척도상으로 가장 우수한 모형의 결과도 함께 보기로 한다.

〈표 10〉 모형의 비교

| 모형 | 선택된 변수 | R^2 | Adj R^2 | AIC |
|----|--|---------------|---------------|----------------|
| 1 | X_0, \dots, X_{18} | 0.2933 | 0.2805 | 5973.51 |
| 2 | $X_0, X_1, X_2, X_3, X_4, X_5, X_7, X_8, X_{11}, X_{13}, X_{15}, X_{18}$ | 0.2915 | 0.2836 | 5962.28 |
| 3 | $X_0, X_1, X_2, X_3, X_4, X_5, X_7, X_8, X_{11}, X_{13}, X_{15}, X_{16}, X_{18}$ | 0.2926 | 0.2840 | 5962.68 |
| 4 | $X_0, X_1, X_2, X_4, X_5, X_7, X_8, X_{11}, X_{13}, X_{15}, X_{18}$ | 0.2888 | 0.2816 | 5963.97 |
| 5 | $X_0, X_1, X_2, X_4, X_5, X_7, X_8, X_{15}$ | 0.2771 | 0.2720 | 5974.35 |

〈표 10〉에서 모형 1은 모든 변수를 선택한 모형으로 R^2 가 가장 높은 모형이다. 이는 다른 모든 모형을 모형 1의 하위모형으로 볼 수 있기 때문에 당연한 결과라 할 수 있다. 즉, 우리는 단순히 독립변수의 수를 늘림으로써 항상 R^2 를 높일 수 있다. 그러므로 연구자들은 R^2 를 모형선택의 척도로 사용하기를 꺼리고 있다. 모형 2는 AIC를 최소로 하는 모형으로 절편을 포함하여 12개의 변수를 선택하였고, 모형 3은 Adjusted R^2 를 극대화하는 모형인데 절편을 포함하여 총 13개의 변수를 가진 모형이다. 변수의 수가 서로 다른 두 모형을 비교할 때 연구자들은 AIC나 Adjusted R^2 를 빈번히 사용하는데 두 척도 모두 R^2 와는 다르게 독립변수 수의 증가에 따라 벌점을 부가한다는 특징을 가지고 있다. 모형 4는 단계식 회귀분석법에 의해 선택된 모형이고 모형 5는 베이지언 변수선택법에 의해 선택된 모형이다.

〈표 10〉으로부터 우리는 베이지언 변수선택법에 의해 선택된 모형이 세 가지 모든 척도에 있어서 가장 나쁜 모델이라는 성급한 결론을 내릴 수 있다. 그러나 베이지언 모델의 독립변수 수가 가장 적기 때문에 R^2 가 가장 낮은 것은 당연한 결과이다. 또한 다른 두 척도 AIC와 Adjusted R^2 역시 모형 선택을 하는데 있어 매우 문제가 있는 척도라는 점을 우리는 강조한다. 바로 이 척도의 불완전성 때문에 단계식 회귀분석법을 포함한 다른 모델이 베이지언 모델에 비해 더 낮은 AIC를 가지거나 더 높은 Adjusted R^2 를 가진다고 할지라도 이를 모형이 더 나은 모형이라고 말할 수는 없는 것이다. 즉, 모형 2나 3에서와 같이 모든 가능한 회귀분석을 하여 가장 낮은 AIC를 가지는 모형이나 가장 높은 Adjusted R^2 를 가지는 모형이 진실한 모형이라는 보장은 없다는 것이다. 단계식 회귀분석법이 두 가지 척도에 있어 베이지언 방법론에 비해 우수한 근본적인 이유는 두 가지 척도 모두 오차 자승의 합 (sum of

squared errors)의 함수이고 단계적 회귀분석법은 변수를 선택할 때 그 기준으로 사용하는 F 값이 오차 자승의 합의 함수이기 때문이다. 반면, 베이지언 변수선택법은 각 변수의 이후 확률분포를 계산하여 모수를 추정하기 때문에 반드시 위의 두 척도를 최소화 또는 최대화하지 않는다. 우리의 이 주장은 모형 선택에 있어 매우 중요한 논제로 올바른 척도 개발을 위한 연구의 필요성이 있음을 보여 주는 부분이라고 하겠다. 올바른 척도가 없는 이 시점에서 우리는 모델의 평가를 위한 가장 합리적인 방법으로 시뮬레이션을 주장하는 바이다.

우리의 주장을 보다 명확히 하기 위해 첫 번째 시뮬레이션과 유사한 실험을 하기로 한다. 모든 가능한 모델의 회귀분석을 수행하기 위해서는 독립변수의 수를 줄여 서로 독립적인 30개의 독립변수를 표준정규분포 $N(0,1)$ 에서 100개씩 추출하기로 한다. 그리고 ϵ 을 독립적으로 표준정규분포 $N(0,1)$ 에서 100개 추출하고 우리는 $Y = \epsilon$ 이라 가정한다. 즉, 진정한 모델은 어떤 독립변수도 포함하지 말아야 한다. 그러나 이 데이터에 가장 낮은 AIC를 가지는 모델은 9개의 독립변수를 가진 모형이었고 가장 높은 Adjusted R^2 를 가지는 모델은 12개의 독립변수를 가지는 모델이었다. 반면 베이지언 변수선택법은 아무 변수도 포함하지 않는 모형이었다. 이 예는 과거의 연구자들이 모델을 선택하는데 보편적으로 사용하는 두 척도의 문제점을 극명하게 보여 주고 있다.

추정샘플 내에서의 모형을 평가하는데 따르는 어려움을 간접적으로나마 극복하고자 우리는 추정하는데 이용하지 않은 398명의 샘플에 위의 다섯 가지 모형을 적용하여 얼마나 예측력이 좋은가를 평가하기로 한다. 우리는 모형의 예측력 또는 오차를 측정하기 위해 두 가지 척도를 쓰기로 하는데 그 첫째는 평균자승오차(Mean squared error)이고 다른 하나는 평균 절대오차(Mean absolute error)이다. 평균자승오차는 $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 / n$ 으로 정의하고 평균절대오차는 $\sum_{i=1}^n |Y_i - \hat{Y}_i| / n$ 으로 정의한다.

〈표 11〉 추정외 샘플에서의 모형의 비교

| 모형 | 평균자승오차 | 평균절대오차 |
|----|--------|--------|
| 1 | 294.6 | 13.28 |
| 2 | 296.8 | 13.34 |
| 3 | 294.2 | 13.26 |
| 4 | 299.9 | 13.38 |
| 5 | 285.6 | 13.15 |

〈표 11〉은 비록 그 정도의 차이는 크지 않지만 두 가지 척도 모두에서 베이지언 변수 선택법에 의해 선택된 모델이 가장 좋은 모형이고 단계적 변수 선택법이 가장 나쁜 방법론으로]

평가 되었다. 또한 베이지언 모델은 가장 단순한 모형으로 그 변수의 수가 다른 모형에 비해 최소 3개 이상 적은 점을 고려할 때 <표 11>의 결과는 대단히 흥미로운 결과이다.

V. 결 론

많은 연구자들은 선형회귀분석에서 일군의 독립변수를 선택할 때 적용의 단순성이라는 이유 하나만으로 단계적 회귀분석법을 빈번히 사용하고 있다. 본 연구는 이와 같은 관행을 비판하면서 보다 발전된 변수 선택 방법론으로 베이지언 변수선택법을 제안하고 있다. 그러나 아직 이를 수행하기 위한 보편적인 통계 소프트웨어가 존재하지 않는 점이 쉽게 베이지언 변수선택법을 이용하는데 있어 장애요인이 될 수 있다.¹⁾ 하나의 대안으로써 우리는 부분 회귀분석법의 사용을 권고한다. SAS, SPSS, BMDP, S-Plus등의 보편적인 통계 팩키지는 독립변수의 수가 20~30개 이하인 경우 부분회귀분석을 수행하여 사용자가 사전에 정한 척도에 -예를 들면 Adjusted R^2 , AIC등등 -의해 가장 좋은 모형을 선택하는 기능을 가지고 있다. 만약 독립변수의 수가 30개를 초과할 경우는 먼저 단계적 회귀분석법을 적용하여 변수의 수를 30개 미만으로 줄인 이후 부분 회귀분석법을 적용하는 것을 권고하는 바이다.

그러나 본 논문은 척도 자체가 가지는 문제점 때문에 부분 회귀분석법을 적용하여 가장 좋은 모형을 선택하는 방법에도 문제가 있다는 점을 지적하고 있다. 이에 반해 본 연구는 데이터 시뮬레이션에 의해 다양한 데이터 구조 하에서 베이지언 변수선택법이 진정한 모형을 발견하는데 우수한 방법론임을 보이고 있다. 또한 실제 마케팅데이터에 베이지언 방법론을 적용하면서 베이지언에 의해 선택된 모형이 다른 방법론에 의해 선택된 모형보다 독립변수의 수는 최소 3개 이상 적은데도 불구하고 예측력은 오히려 더 우수하다는 점을 보이고 있다.

본 연구는 선형회귀모형에서의 변수 선택을 위한 새로운 방법론으로 베이지언 변수선택법을 제시하고 있는데, 이 문제를 다른 각도에서 접근하는 주목할 만한 방법론이 있다. Genetic Algorithm이라 불리는 일종의 데이터마이닝 기법을 변수선택 문제에 적용한 방법론으로, 몇몇 학자들은 변수선택에 있어 Genetic Algorithm이 단계적 회귀분석법보다 뛰어나다고 주장하는 논문을 발표하였다 (Wasserman and Sudjianto 1994; Sudjianto, Wasserman and Sudarbo 1996). 한가지 흥미로운 향후 연구과제로 베이지언 변수선택법

1) 본 연구의 베이지언 변수선택법을 수행하기 위해 필자는 Smith M.이 제공한 code를 기초로 하여 Fortran프로그램을 코딩하였다. 필자는 관심이 있는 연구자에게 무료로 이 Fortran 프로그램을 공급하고 있다.

과 Genetic Algorithm을 비교하는 연구를 제언한다.

본 연구의 주제인 독립변수 선택의 문제는 사실 통계학의 커다란 연구 분야의 하나인 “모형 선택의 문제”의 일부로 볼 수 있다. 즉, 종속변수의 변환 방법에 따라 우리는 로짓모델(logit model), 프로빗모델(probit model) 등의 모형을 도출할 수 있는데 데이터로부터 변환방법을 선택하도록 하는 기법이 모형선택 문제의 하나의 예이다. 또한 각 독립변수를 그대로 회귀분석 모형에 적용할 것인지 아니면 어떤 변수 변환을 하여야 하는 것인지의 선택의 문제 역시 모형선택의 문제이다. 우리는 이 모형선택의 문제를 또 하나의 향후 연구과제로 제언한다.

참 고 문 헌

- Berk, K. (1978), "Comparing Subset Regression Procedures," *Technometrics*, 20, 1, 1-6.
- Casella, G., and George, E. (1992), "Explaining the Gibbs Sampler," *The American Statistician*, 46, 167-174.
- Draper, N. and Smith, H. (1981), *Applied Regression Analysis*, 2nd edition, New York: John Wiley & Sons, Inc.
- Edwards, D. and Havranek, T. (1987), "A Fast Model Selection Procedure for Large Families of Models," *Journal of the American Statistical Association*, 82, 205-213.
- Furnival, G. and Wilson, R. (1974), "Regression by Leaps and Bounds," *Technometrics*, 16, 4, 499-511.
- Gelfand, A. and Smith A. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398-409.
- George, E. and McCulloch R. (1993), "Variable Selection Via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 423, 881-889.
- George, E. and McCulloch R. (1996), "Approaches For Bayesian Variable Selection," *Working Paper*, The University of Chicago.
- Hocking, R. (1976), "The Analysis and Selection of Variables in Linear Regression," *Biometrics*, 32, 1-49.

- Kim, B. and Park, K. (1997), "Studying Patterns of Consumer's Grocery Shopping Trips," *Journal of Retailing*, 73, 4, 501-517.
- Madigan, D. and Raftery, A. (1994), "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window," *Journal of the American Statistical Association*, 89, 428, 1535-1546.
- Miller, A. (1984), "Selection of Subsets and Regression Variables (with discussion)," *Journal of Royal Statistical Society, A*, 147, 389-429.
- Raftery, A., D. Madigan, and Hoeting, J. (1993), "Model Selection and Accounting for Model Uncertainty in Linear Regression Models," *Working Paper*, Department of Statistics, University of Washington.
- Smith, M. and Kohn, R. (1996), "Nonparametric Regression Using Bayesian Variable Selection," *Journal of Econometrics*, 75, 317-343.
- Sudjianto, A., G. Wasserman, and Sudarbo, H. (1996), "Genetic Subsets Regression," *Computers & Industrial Engineering*, 30, 4, 839-849.
- Tanner, M. and Wong W. (1987), "The Calculation of Posterior Distributions by Data Augmentation" (with discussion), *Journal of the American Statistical Association*, 82, 528-550.
- Wasserman, G. and Sudjianto, A. (1994), "All Subsets Regression Using a Genetic Search Algorithm," *Computers & Industrial Engineering*, 27, 489-492.
- Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, New York: John Wiley & Sons, Inc.