응용통계연구 제 14권 2호, 2001년, pp. 321-332

모의 담금질을 이용한 이진반응변수 사영추적회귀 *

박종선[1)]

요 약

본 논문에서는 반응변수가 두 가지의 값을 갖는 회귀분석에 적용할 수 있는 사영추적회귀를 고려하였다. 회귀모형에 필요한 설명변수들의 선형결합이 하나이고 연결함수의 형태를 사전에 알지 못한다는 가정하에서 모의담금질 기법을 이용하여 모형에 필요한 선형결합을 찾는 알고리즘을 제시하였다. 이진 반응변수의 경우에는 평활모수의 값에 따라 잔차이탈도함수의 반응표면이 단봉의 형태를 갖지 않는 경우가 있어 비동질적 마코프체인을 이용한 모의담금질 기법을 적용하면 효율적으로 선형결합을 탐색할 수 있다.

주요용어: 사영추적회귀, 이진반응변수, 모의담금질.

1. 서론

크기가 n인 반응변수의 벡터를 $y = \{y_i\}$라 하고 p차원의 설명변수들을 $n \times p$차원의 행렬인 $X = (x_1, \ldots, x_p)$라고 하자. 이 때 반응변수는 "성공" 또는 "실패", "사망" 또는 "생존" 등과 같이 두 가지의 경우만을 생각하고 값으로 0 또는 1을 적당히 주었다고 가정하자. 이 경우 회귀분석은 설명변수들이 주어졌을 때 반응변수의 조건부 분포에 관한 모든 내용을 포함한다고 할 수 있다. 그러나 현실적으로는 조건부 평균 $E(y|X)$ 과 분산 $Var(y|X)$ 등에 대한 특성을 파악하는 데 중점을 두고 있다.

일반적인 회귀분석은 사전에 가정된 회귀함수 또는 회귀표면(regression surface)을 이용하여 미지의 회귀계수들을 추정하게 된다. 그러나 설정된 모형이 참이 아닌 경우에는 도출된 결과들을 신뢰할 수 없게 되며 모형이 참인 지 아닌 지를 판단하는 것 또한 쉬운 일이 아니다. 이와 같은 이유로 p-차원의 국소 평균법을 이용하는 비모수적 회귀모형에 대한 연구가 많이 있어 왔으나 표본의 수가 설명변수의 차원에 비하여 상대적으로 적은 경우에는 잘 알려진 희박성문제(sparseness problem) 때문에 사용에 제한을 받아 왔다.

Friedman과 Stuetzle (1981)은 이와 같은 문제를 해결하는 하나의 방법으로 각각의 회귀계수를 추정하는 대신 회귀모형에 필요한 설명변수들의 선형결합들을 찾는 사영추적회귀 (projection pursuit regression)를 제시하였다. 이 방법의 장점 중 하나는 회귀함수의 형태에 대한 가정이 없이 회귀모형에 필요한 선형결합들을 찾는다는 점이다.

이 논문에서는 이진 반응변수를 갖는 회귀모형에 사영추적회귀방법을 적용하여 보았다. 일반화 선형모형에서와 같이 설명변수들이 주어졌을 때 반응변수의 조건부 분포는 이항분

* 이 논문은 1998년도 성균관대학교 석천연구비에 의해 연구되었음.

1) (110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 경제학부 통계학전공, 부교수
E-mail: cspark@skku.ac.kr

포를 가정하였다. 따라서 목적함수도 잔차제곱합이 아닌 잔차이탈도함수(residual deviance function)가 사용되었으며 연결함수에 대한 가정은 하지 않고 평활이 가능한 함수이면서 0과 1사이의 값을 갖는 함수를 가정하였다. 그리고 회귀모형에 필요한 설명변수들의 선형결합이 하나라는 가정하에서 비동질적(inhomogeneous) 마코프체인을 이용한 모의담금질(Simulated Annealing)기법을 적용하여 이를 찾는 방법을 제시하였다.

본 논문의 구성은 다음과 같다. 먼저 제 2 장에서는 본 논문에서 사용될 모형을 가정하였으며, 제 3 장에서 회귀함수에 필요한 선형결합을 모의담금질 방법을 이용하여 찾는 알고리즘을 제시하였다. 제 4 장과 제 5 장에서는 모의자료와 실제자료를 통하여 새로운 기법의 활용가능성을 보였으며 마지막으로 결론을 제 6 장에 포함하였다.

2. 모형 및 평활함수에 대한 가정

설정된 모형은 다음과 같다. 우선 반응변수 y는 다음과 같은 평균을 갖는 베르누이 분포를 따른다고 가정한다.

$$E(y|X) = f(\eta) = S\left(\frac{\exp(\eta)}{1+\exp(\eta)}\right) \tag{2.1}$$

이 때 S는 평활함수이고 $\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ 즉, 설명변수들의 선형결합이다. 여기서 평활함수 S의 모수를 단순히 η로 하지 않고 로지스틱(logistic)함수를 사용한 이유는 반응변수의 추정치들이 0과 1사이의 값을 갖도록 하기위한 것이다. 다시 말하면 logistic(η)를 새로운 설명변수라고 생각할 수 있으며 로지스틱함수와 평활함수의 결합함수 f가 바로 회귀모형에 필요한 참의 연결함수가 된다. f는 미지의 회귀함수라고 가정하자.

만일 평활모수의 값에 관계 없이 반응변수의 조건부 분포가 이항분포라는 가정하에 대수우도함수의 곡면이 단봉의 형태를 갖는 매끄러운 곡면일 경우에는 기존의 사영추적 알고리즘을 사용하면 대수우도함수를 최적으로 하는 설명변수들의 선형결합을 찾을 수 있다. 그러나 이항분포의 경우 대수우도함수의 곡면은 평활모수의 값에 따라 하나 이상의 국소최소값을 갖는 경우가 생기게 된다. 이것은 반응변수가 단 두 가지의 값을 갖는 특성 때문으로 생각되며 따라서 평활모수의 값에 따라 대수우도함수의 곡면은 여러 가지의 형태를 가지게 된다. 이를 자세히 살펴보면 그림 2.1, 그림 2.2, 그리고 그림 2.3과 같다. 전체적으로 평활모수가 작은 값을 갖는 경우에는 국소 최소값이 여러 개가 되며 큰 경우에는 곡면이 너무 완만하여 전체적 최적값을 찾지 못하거나 정확한 값을 찾지 못하게 되는 경우가 생길 수 있음을 볼 수 있다. 그림들은 4.1절의 모의자료 1을 사용하여 작성되었다.

본 논문에서는 사영추적회귀방법에서 사용되는 수퍼평활기(super smoother)를 이용하는 대신 모의담금질기법을 적용할 때 평활모수인 밴드폭(bandwidth)을 임의로 지정할 수 있는 가우시안 함수를 이용한 핵(kernel)평활법을 사용하는 경우에 대하여만 언급하기로 한다. 그 이외의 다른 평활법, 예를 들면 평균평활법이나 로에스(loess) 방법 등도 결과 값이 0과 1 사이의 값을 갖고 평활모수인 밴드폭을 임의로 지정할 수 있다면 아무런 문제가 되지 않는다.

그림 2.1: 왼쪽:평활모수 0.1 오른쪽:평활모수 0.2 인 경우의 잔차이탈도

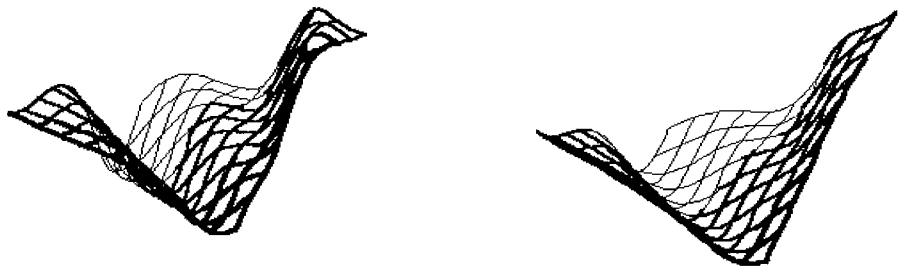

그림 2.2: 왼쪽:평활모수 0.3 오른쪽:평활모수 0.4 인 경우의 잔차이탈도

그림 2.3: 왼쪽:평활모수 0.6 오른쪽:평활모수 0.9 인 경우의 잔차이탈도

3. 탐색 알고리즘

최적화를 위한 목적함수인 잔차이탈도함수를 구하기 위해 먼저 대수우도함수를 살펴보면 다음과 같다.

$$l(y,\ \beta) = \Sigma_{i=1}^{n}\{y_i \log p_i + (1-y_i)\log(1-p_i)\} \tag{3.1}$$

여기서 $p_i = E(y_i|X_i) = f(\eta_i)$가 된다. 물론 앞에서 언급한 것처럼 함수 f는 0과 1사이의 실수값을 가지며 평활함수 S와 로지스틱함수의 복합함수 (composite function) 이다. 이제 목적함수인 잔차이탈도함수는 다음과 같다.

$$\begin{aligned} D(y,\ \beta) &= 2l(y,\ \tilde{\beta}) - 2l(y,\ \beta) \\ &= -2\Sigma\{y_i \log(\frac{p_i}{1-p_i}) + \log(1-p_i)\} \end{aligned} \tag{3.2}$$

잔차이탈도함수는 가능한 대수우도함수의 최대값($\tilde{\beta} = y$에서 얻어진다고 가정)과 대수우도함수의 차이에 2를 곱한 값이다. 목적함수는 β값의 변화에 따라 η값이 변화하고 f함수를 통하여 p_i를 결정하게 된다. 이 때 식 (3.2)의 잔차이탈도함수 D을 극소화하는 β를 찾는 것이 목적이다.

3.1. 모의담금질 (Simulated Annealing) 기법

Kirkpatrick, Gelatt, 그리고 Vecchi (1983)에 의하여 소개된 모의담금질 기법은 일반적으로 여러 개의 국소최적값이 있는 경우 전체최적을 찾는 문제에 적용되어 왔다. 특히 조합최적화문제(combinatorial optimization problem)의 경우에 효과적으로 해를 찾을 수 있는 방법으로 알려져 있으며 대표적인 예로는 세일즈맨 여행문제를 들 수 있다. 이 방법의 핵심은 최적화하고자 하는 목적함수를 이용하여 분포를 만들고 이 분포에서 마코프체인 몬테칼로 기법을 통하여 표본을 추출하는 것이다. 이러한 표본을 정해진 비율로 선택하면서 동시에 온도라 불리는 값이 낮추어가면 선택된 표본들이 최적해에 수렴하는 것을 이용하여 원하는 해를 구하게 된다.

본 문제에서 목적함수는 잔차이탈도 $D(y,\ \beta)$이므로 이를 이용하여 만들어진 분포는 다음과 같다.

$$u(\beta) = C\ \exp\left(\frac{1}{\gamma}D(y,\ \beta)\right) \tag{3.3}$$

여기서 C는 상수항이며 분포의 확률변수는 β가 된다. 이제 보통 "온도"라 불리는 γ값을 서서히 낮추어가면서 각 단계에서 β에 대한 표본을 구하게 되고 정해진 비율로 표본을 선택하게 되면 표본은 목적함수를 최적화하는 점으로 수렴하게 된다. 구체적인 알고리즘은 다음과 같다.

(1단계) 초기화

(a) $U(-1,1)$에서 초기 β 지정

(b) $k=0$ (step function)

(c) 초기온도 γ_0 설정 ($k=0$에서의 온도)

(d) 반복수 L_0 초기화 ($k=0$에서의 반복(표본추출)수)

(e) 초기 평활모수 설정

(2단계) 반복

(a) $l=1$에서 L_k까지

i. 이전 β_i의 주위(neighbor)에서 새로운 β_j 설정

ii. $l(\beta_j) \le l(\beta_i)$이면 $\beta_i = \beta_j$로 치환

iii. $l(\beta_j) > l(\beta_i)$이고 $\exp(\frac{l(\beta_i)-l(\beta_j)}{c_k}) > U[0,1]$이면 $\beta_i = \beta_j$로 치환

(b) $k = k+1$로 치환

(c) k 단계에서의 평활모수 설정

각 단계 k에서 평활모수의 값 또한 0과 1사이의 작은 값에서 큰 값으로 적당히 변화시킨다. 간단한 시뮬레이션의 결과 평활모수는 대체적으로 $0.1 \sim 0.2$의 값에서 시작해서 단계별로 $0.01 \sim 0.05$정도씩 증가 시키면 별 무리가 없는 것으로 나타났다.

이처럼 k값에 따라 변화하는 평활모수와 목적함수를 가지고 만들어지는 분포에서의 표본추출은 비동질적 (inhomogeneous) 마코프체인을 이용하게 되고 이러한 경우에 대한 모의담금질 기법의 수렴에 대한 연구는 Mitra와 동료들 (1986), Anily와 Federgruen (1987)에 자세히 연구되어 있다.

3.2. 해의 수렴확인을 위한 그래픽스

앞의 절에서 언급한 것처럼 평활모수의 값이 작은 경우에는 목적함수의 표면이 여러 개의 국소 최대값을 갖고 평활모수의 값이 큰 경우에는 목적함수의 표면이 너무 완만하게 된다. 따라서 최적해는 적당한 평활모수의 값에서 구해진 해가 될 것이다. 구해진 해가 최적해임을 확인하기 위해서는 해의 변화 상태와 목적함수값의 변화상태를 동시에 고려하여야 한다. 우리는 이를 관찰하기 위하여 두 가지의 플롯을 제시하였다. 첫 번째 플롯은 해의 변화를 추적하기 위한 것으로 과거의 해에 대하여 새로운 해의 표준화된 거리(standardized distance)를 표시하였다. 그리고 두 번째 플롯에서는 목적함수 즉, 잔차이탈도함수의 값을 플롯하였다.

두 플롯의 특성을 살펴보면 사영추적회귀의 특성상 최적해는 각 회귀계수의 추정치가 아닌 회귀계수의 방향으로 나타나므로 구해진 해들이 최적해와 차이가 많은 초기상태에는 해의 방향(첫번 째 플롯)과 목적함수의 값에 많은 변화가 있으며 해가 일단 최적해에 가까워지면 해의 방향에 변화가 거의 없고 목적함수 또한 기존의 해에 비하여 작은 값을 갖고 변화하지 않게 된다. 여기서 계속 표본을 구하게 되면 평활모수는 계속 증가하게 되어 목

적함수의 표면이 완만하여지고 이 경우 목적함수의 값에 변화가 거의 없는 상태에서 해의 방향이 갑자기 변화하는 현상이 발생하게 되며 이는 두 그래프를 통하여 쉽게 발견할 수 있다. 결론적으로 이러한 상태가 나타나기 직전의 해가 최적해일 가능성이 가장 높다고 할 수 있다. 구체적인 플롯들과 이에 대한 설명은 다음의 모의자료와 실제자료분석에서 살펴보기로 한다.

본 논문에서 제시한 방법에 대한 프로그램은 Xlisp-Stat (Tierney, 1990)을 이용하여 작성되었으며 일반화선형모형과 기존의 사영추적회귀모형의 적합은 S-Plus에서 제공하는 함수를 사용하였다.

4. 모의자료

이 장에서는 모의자료를 이용하여 본 논문에서 제시한 방법을 통하여 최적해에 접근하는 과정을 살펴보고 기존의 사영추적회귀 및 일반화선형모형의 해와 비교해 보기로 한다. 모든 모의자료 및 실제자료의 적합에서 사용된 핵함수(kernel function)는 가우시안 함수이다.

4.1. 모의자료 1

모의자료1에서는 연결함수가 로지스틱함수이고 설명변수는 2개로 각 각 독립인 표준정규분포에서 무작위로 추출되었다. 즉, $\{x_1,\ x_2\} \sim \mathrm{N}_2(0,\ I)$ 이고 표본은 다음 식

$$\mathrm{logit}(p) = x_1 + 2x_2$$

을 만족하는 p를 모수로 갖는 베르누이 분포에서 무작위로 추출되었으며 표본의 수 n은 100이다. 이 자료에 일반화선형모형(GLM)과 사영추적회귀(PPR)를 적합시킨 결과와 본 논문의 방법(SA)을 적용하여 얻어진 회귀계수의 추정치들은 다음의 표 4.1과 같다.

표 4.1: 모의자료 1에 대한 회귀계수 추정치

	GLM (logistic)	PPR	SA
β_1	1.098	0.471	-0.496
β_2	2.058	0.882	-0.752

로지스틱연결함수를 갖는 일반화선형모형, 사영추적회귀, 그리고 모의담금질을 이용한 세 방법에 따른 결과는 일반화선형모형의 경우 두 계수의 추정치 비가 2에서 크게 벗어나지 않았으나 사영추적회귀와 모의담금질의 경우에는 계수의 비가 각 각 1.67과 1.51로 2보다 작은 것으로 나타났다.

참고로 모의자료 1에서 사용된 방법에 대하여 구체적으로 살펴보면 β의 초기 추정치는 앞에서 언급한 데로 $U(-1,\ 1)$에서 무작위로 선택되었으며 새로운 값은 기존의 값에 무작위로 선택된 $U(-1,\ 1)$에 0.01을 곱한 값을 더하여 만들었다. 그 이외에 다른 모수들의 값은

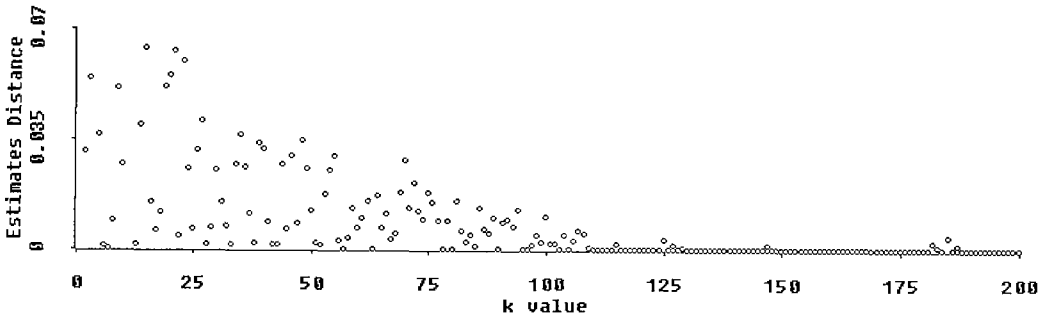

그림 4.1: k값의 변화에 따른 해의 방향(모의자료 1)

다음과 같다. 먼저 초기단계($k = 0$)에서 초기온도 $\gamma_0 = 1000$으로, 초기평활모수는 0.2로 설정하였다. 그리고 k번째 단계에서

- 온도 $\gamma_k = \gamma_0 \times (0.9)^k$
- 반복수 $L_k = 100 + 10 \times k$
- 평활모수는 $0.2 + 0.003k$

이다.

해의 수렴과정을 확인하기 위한 앞 절에서 언급한 두 개의 플롯을 통하여 결과를 살펴보면 그림 4.1 및 그림 4.2와 같다. 두 플롯을 통하여 k의 값이 약 180을 넘어서면 해의 방향에 거의 변화가 없어지며 잔차이탈도 또한 평활모수의 증가에 따라 자연적으로 약간씩 증가하는 것을 제외하면 거의 변화가 없이 수렴하는 것을 볼 수 있다. 특히 두 번째의 잔차이탈도값에 대한 플롯에서는 이 값의 증감 횟수를 통하여 평활모수의 변화에 따른 잔차이탈도함수의 표면형태를 대략적으로 파악할 수 있고 이러한 정보를 바탕으로 k값의 변화에 따른 평활모수의 초기값, 증가분, 그리고 최종값 등을 결정할 수 있다.

4.2. 모의자료 2

모의자료 2에서는 모의자료 1에서 사용한 것과 동일한 설명변수, 연결함수, 그리고 계수들을 사용하였다. 다른 모수들도 모의자료 1과 같다. 다만 모형에 필요하지 않은 설명변수를 추가한 후 결과를 살펴보았다. 추가된 설명변수는 $\{x_3, \ldots, x_{10}\} \sim \mathrm{N}_8(0, I)$을 사용하였다. 각 각의 방법에 따른 해를 살펴보면 다음과 같다.

이 경우에도 세 방법 모두 비슷한 결과를 보였는데 일반화선형모형의 경우 β_9와 β_{10}이 사영추적회귀의 경우 β_{10}이 그리고 모의담금질 방법의 경우 β_9의 추정치가 $\beta_3 - \beta_{10}$의 추정치들에 비하여 상대적으로 0에서 멀리 떨어진 값을 나타내었다.

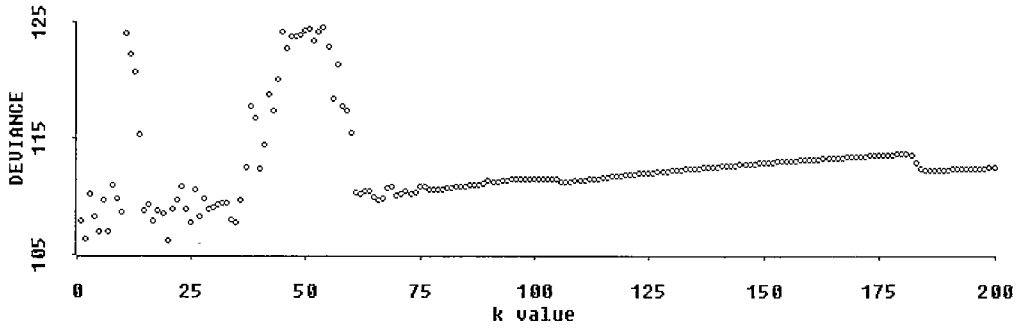

그림 4.2: k값의 변화에 따른 잔차이탈도의 변화(모의자료 1)

표 4.2: 모의자료 2에 대한 회귀계수 추정치

	GLM (logistic)	PPR	SA
β_1	1.344	0.459	0.455
β_2	2.205	0.818	0.999
β_3	0.228	0.070	0.107
β_4	-0.050	0.037	0.057
β_5	-0.133	-0.040	-0.004
β_6	0.131	0.065	0.140
β_7	-0.323	-0.136	-0.184
β_8	0.103	-0.003	0.029
β_9	-0.401	-0.097	-0.361
β_{10}	-0.426	-0.283	-0.081

4.3. 모의자료 3 및 4

모의자료 3에서는 연결함수가 프로빗(probit)이고 설명변수는 모의자료 1에서 사용된 것과 동일한 것을 사용하였다. 다른 모수도 모의자료 1과 같다. 그리고 모의자료 4에서는 모의자료 2에서 사용된 것과 동일한 설명변수를 사용하였으며 연결함수가 프로빗이라는 점만 상이하며 모형은 모의자료2와 동일하다. 결과를 살펴보면 표 4.3 및 표 4.4와 같다.

먼저 모의자료 3에 대하여 살펴보면 일반화선형모형의 경우는 프로빗 및 로지스틱 연결함수의 경우, 그리고 모의담금질 방법의 경우 모두 β_2가 β_1의 2.5배 정도의 크기로 나타났으나 사영추적회귀방법에서는 두 계수 추정치의 비율이 3에 가깝게 나타나 기존의 사영추적회귀가 이산형 등의 반응변수자료에 적용될 경우 문제가 있을 수도 있음을 알 수 있다.

표 4.3: 모의자료 3에 대한 회귀계수 추정치

	GLM (probit)	GLM(logistic)	PPR	SA
β_1	0.817	1.396	0.344	0.373
β_2	1.978	3.324	0.939	0.999

표 4.4: 모의자료 4에 대한 회귀계수 추정치

	GLM (probit)	GLM(logit)	PPR	SA
β_1	1.022	1.797	0.344	0.390
β_2	2.583	4.482	0.922	0.999
β_3	-0.063	-0.116	0.041	0.002
β_4	0.073	0.140	0.008	0.010
β_5	-0.078	-0.128	0.022	-0.023
β_6	0.150	0.260	0.103	0.080
β_7	0.082	0.178	-0.016	-0.016
β_8	0.060	0.099	0.050	0.059
β_9	-0.589	-1.064	-0.120	-0.219
β_{10}	0.177	0.279	0.043	0.070

모의자료 4에서 각 각의 모형에 따른 두 계수 β_1, β_2의 비율은 모의자료3의 결과와 비슷하나 일반화선형모형의 경우에는 연결함수의 형태에 관계없이 β_9의 추정치가 큰 음의 값을 나타내 상대적으로 다른 두 방법에 비하여 정확하지 않은 결과를 나타내었다.

5. 실제 자료

실제분석에서 사용된 자료는 남아프리카 공화국 Western Cape의 세 지역에서 심장동맥 고혈압원인연구를 위하여 사용된 서베이자료(Rousseauw와 동료들, 1983)의 일부이며 Heart-attack자료로 잘 알려져 있다. 서베이의 목적은 위의 지역에서 높은 발병율을 보이고 있는 Ichaemic 심장병의 원인을 규명하기 위한 것으로 자료는 15세에서 64세 사이의 백인을 대상으로 조사되었으며 반응변수는 서베이 당시 myocardial infarction (MI)이 존재하는지 아닌 지에 따라 0과 1의 값을 주었다. 전체 조사자 462명 중 162명은 MI가 존재하였으며 302명은 존재하지 않았다. 위험요인 (risk factor)으로 설정된 설명변수는 혈압(systolic blood pressure - sbp)과 콜레스테롤비(lbl = (전체콜레스테롤 - HDL)/HDL)가 사용되었다. 표 5.1의 결과에서 두 회귀계수 추정치의 비를 살펴보면 로짓 연결함수를 사용한 일반화선형모형의 경우와 모의담금질 방법의 결과가 비슷하였으며 나머지 두 방법의 경우가 서로 비슷한 결과를 나타내었다.

표 5.1: 실제자료에 대한 회귀계수 추정치

	GLM (logistic)	GLM(probit)	PPR	SA
sbp	0.017	0.010	0.067	0.010
lbl	0.255	0.155	0.998	0.184

6. 결론

우리는 앞에서 회귀분석의 반응변수가 이진값을 갖는 경우 모의담금질을 이용한 사영추적회귀를 통하여 모형에 필요한 설명변수들의 선형결합을 찾는 방법에 대하여 살펴보았다. 일반적인 일반화선형모형들은 연결함수에 대한 가정이 필요하기 때문에 가정된 연결함수가 사실과 다른 경우 일반적으로 계수 추정치 등이 일치성을 만족하지 않는 것으로 알려져 있다. 사영추적회귀의 경우 연결함수에 대한 가정은 필요하지 않지만 반응변수가 연속형이 아닌 경우에 대한 적용사례나 이에 대한 연구는 많지 않아 반응변수가 이진값을 갖는 경우에 대한 무조건적인 적용은 잘못된 결과를 얻을 가능성이 있다. 이는 위의 몇 가지 예에서 보듯이 평활모수의 값에 따라 잔차이탈도의 반응표면이 단봉을 갖지 않는 경우를 보아도 알 수 있으며 따라서 이를 고려하지 않은 사영추적회귀의 적용은 잘못된 결과를 가져올 것이다.

본 논문에서 제시한 방법은 일반화선형모형과 같은 모수적 방법들이 연결함수에 대한 가정을 필요로 하는 단점과 사영추적회귀에서 정해진 평활모수의 값에서 잔차이탈도의 반응표면이 단봉의 형태를 갖지 않는 경우 구해진 해는 국소최적에 빠질 수 있다는 점을 보완하는 대안으로 제시되었다. 앞에서 살펴본 몇 가지의 모의자료와 실제자료의 결과에서 우리가 제시한 방법의 우월성을 보일 수는 없었다. 제시된 방법과 기존의 방법간의 비교를 위하여는 더 많은 모의실험이 필요하다고 하겠다. 그러나 최소한 본 논문에서 제시한 방법은 앞에서 설명한 해의 수렴을 확인하기 위한 두개의 그래프와 같이 사용될 경우 국소최적에 빠질 위험이 없으며 임의로 연결함수의 형태를 가정할 필요가 없이 모형에 필요한 설명변수들의 선형결합을 근사적으로 구할 수 있다는 점에서 이진변수와 같이 연속형이 아닌 반응변수들을 갖는 문제들에 적용할 수 있을 것이다.

마지막으로 일반적인 사용추적회귀에서는 1개 이상의 선형결합이 모형에 필요한 경우를 포함하는데 이진반응변수 모형에서는 반응변수와 설명변수의 관계를 정의하는 대신 반응변수의 조건부평균과 설명변수들의 관계를 가정하기 때문에 하나 이상의 선형결합이 포함되는 경우 기존의 사영추적회귀에서와 같이 가법잔차(additive error)를 고려하여 반응변수에서 선형결합의 적당한 평활값을 제한 나머지를 다음의 단계에서 반응변수 값으로 사용할 수가 없었다.

참고문헌

[1] Anily, S. and Federgruen, A. (1987). Simulated annealing methods with general acceptance probabilities, *J. Appl. Probab.*, **24**, 657-667.

[2] Friedman, J.H. and Stuetzle, W. (1981). Projection Pursuit Regression, *Journal of the American Statistical Association,* **76,** 817-823.

[3] Kirkpatrick, S., Gelatt, C.D. Jr. and Vecchi, M.P. (1983). Optimization by Simulated Annealing, *Science*, **220**, 671-680.

[4] Mitra, D., Romeo, F. and Sangiovanni-Vincentelli, A. (1986). Convergence and fintite-time behavior of simulated annealing, *Adv. Appl. Probab.*, **18**, 747-771.

[5] Rousseauw, J., du Plessis, J., Benade, A., Jordaan, P., Kotze, J., Jooste, P. and Ferreira, J. (1983). Coronary risk factor screening in three rural communities, *South African medical journal*, **64**, 430-436.

[6] Tierney, L. (1990). *Lisp-Stat: An Object-oriented Environment for Statistical Computing and Dynamic Graphics*, J. Wiley & Sons, New York.

[7] Tierney, L. (1994). Markov Chains for Exploring Posterior Distributions, *Annals of Statistics*, **43**, 159-78.

[2000년 1월 접수, 2001년 5월 채택]

Projection Pursuit Regression for Binary Responses using Simulated Annealing*

Chongsun Park [1)]

ABSTRACT

In this talk we will propose a projection pursuit algorithm to find relevant linear combinations of predictors in the regression problem with binary responses using simulated annealing. In some cases with binary responses it can be shown that surface of the residual deviance for the mean function is heavily depending on smoothing parameters and even not unimodal. We suggest a method using simulated annealing algorithm to avoid trapping in the local optimum. A graphical tools to access convergence state will also be suggested.

Keywords: Projection Pursuit Regression; Binary Responses; Simulated Annealing.

* This paper was supported by '98 Suk Chun Research Fund, Sungkyunkwan University.

1) Associate Professor, Department of Statistics, Sungkyunkwan University.
E-mail: cspark@skku.ac.kr