응용통계연구 제 **20**권 1호, 2007년, pp. 133–145

국소선형 준가능도 추정량의 자료 희박성 문제 해결방안*

박동련[1)]

요 약

국소선형 추정량은 여러 면에서 바람직한 특성을 많이 갖고 있는 좋은 추정량이다. 그러나 자료가 희박한 부분에서는 매우 불안정한 추정값을 갖게 되는 문제가 있음이 밝혀졌으며, 이 문제를 해결하기 위한 여러 방안이 많이 연구되었다. 그러나 이항반응변수를 위한 국소선형 추정량의 변형이라고 할 수 있는 국소선형 준가능도 추정량에 대해서는 아직 자료의 희박성 문제가 다루어지지 않고 있었다. 이 논문에서는 국소선형 준가능도 추정량이 갖고 있는 자료의 희박성 문제를 인식하고, 몇 가지 해결방안을 제시하였으며, 모의실험을 통하여 가장 효과적인 방안을 선택하였다.

주요용어: 국소선형 준가능도 추정량, 이항반응변수, 유사자료.

1. 서론

이항반응변수는 매우 다양한 분야에서 흔히 접할 수 있는 반응변수의 형태이다. 예를 들어, 기업부실 예측모형에서 반응변수는 '부실기업' 또는 '정상기업'의 두 가지 범주에 속하게 된다. 또한 신상품의 구매의사 성향분석에서 반응변수는 '구매' 또는 '구매보류'의 두 범주로 구분된다. 이러한 경우 반응변수의 i번째 관찰값 Y_i는 $Y_i = 1$ 또는 $Y_i = 0$의 값을 부여 받게 된다. 이항반응변수의 분석에서 하나의 설명변수 X만을 고려하고 있는 경우에 우리는 다음과 같은 가정을 하게 된다.

$$P(Y_i = 1|X_i = x_i) = p(x_i) = 1 - P(Y_i = 0|X_i = x_i), \quad i = 1, \ldots, n. \tag{1.1}$$

여기에서 함수 $p(\cdot)$를 반응곡선이라고 하는데, 반응곡선의 추정방법은 크게 모수적인 방법과 비모수적인 방법으로 구분된다. 비모수적인 추정방법으로는 Müller와 Schmitt(1988)에서 사용된 커널 추정량이나 Park(1999)에서 사용된 국소선형회귀모형을 사용할 수 있다.

그러나 이러한 추정량들은 모두 이항반응변수가 갖고 있는 특성을 제대로 살릴 수 없기 때문에 $P(Y = 1|X = x)$의 추정량으로서는 한계를 가지게 된다. 명확한 문제 중의 하나는 추정된 반응곡선 $\hat{p}(x)$이 0보다 적거나 1보다 큰 값을 가질 수도 있다는 점이다. 이러한 문제는 Fan 등(1995)이 제안한 국소선형 준가능도(quasi-likelihood) 추정량을 사용함으로써 해결할 수 있다. Park과 Park(2006)은 $p^{-1}(\alpha)$의 추정방법으로 국소선형 준가능도 추정량과

* 이 논문은 2007년도 한신대학교 학술연구비 지원에 의하여 연구되었음
1) (447-791) 경기도 오산시 양산동 411, 한신대학교 정보통계학과, 교수
E-mail: drpark@hs.ac.kr

로짓모형에 의한 추정방법의 효율성을 모의실험을 통하여 비교하였고, 비교결과 비모수적인 방법이 상당히 효과적인 방법임이 밝혀졌다.

국소선형 추정량은 비록 많은 장점을 갖고 있는 좋은 추정량이기는 하지만 그 나름대로의 문제점도 갖고 있는 추정량이다. Seifert와 Gasser(1996)는 국소선형 추정량을 실제자료에 적용시켰을 때 기대했던 것보다 훨씬 나쁜 추정결과가 종종 나오게 되는 현상을 지적하였다. 이것은 결국 추정량의 분산이 무한정 큰 값을 갖게 되는 현상을 의미하는 것인데, Seifert와 Gasser는 이 문제가 자료의 희박성(sparsity) 때문에 발생하는 것임을 밝혔다. 이 문제를 해결하는 방법으로 그들은 국소선형추정량을 정의하는 과정에 능형모수를 포함시키는 방법을 제시하였다. 그러나 이 방법은 상대적으로 복잡할뿐더러 능형모수를 선택해야 되는 문제를 안고 있게 된다. 또한 Seifert와 Gasser는 자료가 희박한 구간에서는 띠폭을 늘려 충분한 자료를 이용하게 함으로써 자료의 희박성 문제를 해결하는 방법도 제시하였는데 이 방법은 실질적으로 최근접이웃방법(nearest neighbor)으로 띠폭을 선택하는 방법과 큰 차이가 없다고 하겠다. 그러나 Jennen-Steinmetz와 Gasser(1988)에 의하면 최근접이웃방법이 매우 간단하게 자료의 희박성 문제를 해결할 수 있는 것으로 보이기는 하지만 회귀함수의 곡률을 전혀 고려하지 않고 띠폭을 결정하기 때문에 최적의 결과를 낳을 수 없다고 한다. Hall과 Turlach(1997)는 자료가 희박한 구간에 유사자료를 포함시킴으로 해서 이 문제가 간단하게 해결될 수 있음을 보였고, 모의실험을 통하여 그들의 방법이 Seifert와 Gasser(1996)의 능형모수를 이용하는 방법보다 더 효과적임을 보였다.

자료의 희박성 문제는 국소선형 준가능도 추정량에서도 동일하게 나타날 수 있는 문제로서 자료가 희박한 부분에서는 매우 엉뚱한 결과를 낳을 수 있게 된다. 그림 1.1은 자료가 희박한 구간에서 발생할 수 있는 국소선형 준가능도 추정량의 문제를 보여주고 있다. 자료의 크기는 $n = 20$이며 자료의 위치는 그림 1.1에 작은 원으로 표시되어 있다. X-축 관찰값은 Uniform(−0.5, 1.5)에서 임의로 추출되었는데 공교롭게 (0.3, 0.6)의 구간에서는 자료가 하나도 추출되지 않았다. Y-축의 이항관찰값은 다음의 반응곡선에 의하여 얻어졌다.

$$p(x) = \frac{\exp\left(-4 + 8x + x^2\right)}{1 + \exp\left(-4 + 8x + x^2\right)}. \tag{1.2}$$

식 (1.2)의 반응곡선은 그림 1.1에 실선으로 표시되어 있다. 주어진 자료에 대하여 고정된 띠폭(bandwidth)을 사용한 국소선형 준가능도에 의한 반응곡선의 추정값을 구하였는데 이때에 사용된 고정된 띠폭은 점근적 MISE(Mean Integrated Squared Error)를 최소화 시키는 띠폭을 사용하였고, 추정 결과는 파선(dashed line)으로 표시되어 있다. 최적 띠폭을 사용했음에도 불구하고 실제 반응곡선과는 너무도 큰 차이를 보이고 있음을 알 수 있다. 이렇듯 경우에 따라서는 상당히 심각한 문제를 야기할 수 있는 자료의 희박성 문제가 이항반응회귀모형에서는 아직 다루어지지 않고 있다.

이 논문에서는 이항반응변수에 대한 국소선형 준가능도 추정량이 갖고 있는 자료의 희박성 문제를 인식하고 해결할 수 있는 방안을 제시하고자 한다. 해결방안은 Hall과 Turlach (1997)에서 제시한 방법, 즉 자료가 희박한 구간에 유사자료를 적절히 배치하는 방법을 사용하고자 한다. 문제는 유사자료의 (x, y) 값을 어떻게 결정하는 것이 가장 최상의 결과를 낳을 수 있겠는가 하는 점이 된다. 이 논문에서는 몇 가지의 해결방안을 제시하고자 하는데

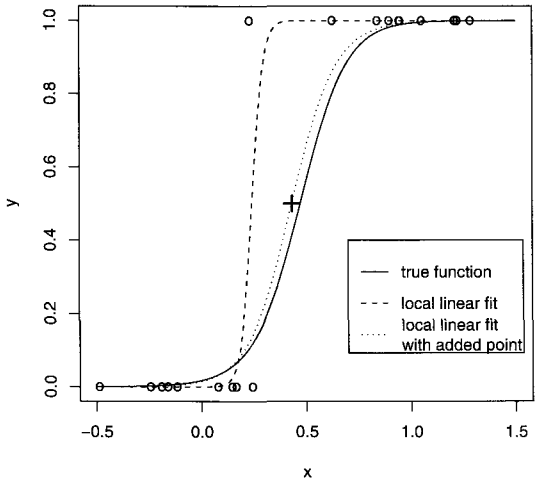

그림 1.1: 자료가 희박한 구간에서 국소선형 준가능도 추정량이 보이고 있는 매우 엉뚱한 결과 및 유사자료 한 개를 추가함으로써 문제가 해결된 것의 예시. (사용된 띠폭은 점근 MISE를 최소화시키는 최적 띠폭으로 $h = 0.83$이 사용되었다.)

그 중 한 해결방안의 결과가 그림 1.1에 나타나있다. 이 방법에 의하면 한 개의 유사자료가 포함되며 그 위치는 십자로 표시된 (0.43, 0.5)가 된다. 점선으로 표시된 추정값은 한 개의 유사자료를 포함한 $n = 21$의 자료에 대한 국소선형 준가능도에 의한 결과이다. 실제 반응곡선에 매우 근접한 결과를 보여주고 있다. 단 한 개의 관찰값을 더 포함시켜 추정했을 뿐인데 그 차이는 너무도 크다는 것을 알 수 있다. 따라서 파선으로 표시된 추정값의 문제는 (0.3, 0.6)의 구간에 자료가 없기 때문에 발생한 것으로 판단할 수 있으며 이러한 자료의 희박성 문제는 유사자료를 적절하게 배치함으로써 해결할 수 있음을 보여주고 있는 것이다.

이 논문에서는 반응곡선의 비모수적 추정량으로 그 효율성이 입증된 국소선형 준가능도 추정량이 갖고 있는 자료의 희박성 문제를 인식하고 해결방안을 제시하도록 하겠다. 2절에서는 국소선형 준가능도 추정량에 대한 대략적인 소개와 자료의 희박성 문제를 인식할 것이고 3절에서는 몇 가지 해결방안을 제시할 것이며 4절에서는 제시된 여러 방법들의 효과를 모의실험을 통하여 비교할 것이다.

2. 국소선형 준가능도 추정량의 자료 희박성 문제

2.1. 국소선형 준가능도 추정량

식 (1.1)에는 하나의 설명변수만을 갖고 있는 이항반응모형이 정의되어 있다. $Y_1, \ldots, Y_n$을 이 모형에서 추출된 서로 독립인 이항반응변수라고 하자. 반응곡선 $p(x)$는 $p \in \mathcal{C}^2([0,1])$라

고 가정하자.

반응곡선 $p(x)$를 모수적 일반화선형모형으로 추정하고자 한다면 다음과 같이 모형화 시킬 수 있다.

$$\eta(x) = \beta_0 + \beta_1 x = g(p(x)). \tag{2.1}$$

단, 함수 g는 연결함수를 의미하며 정준연결함수는 로짓함수가 된다. 로짓모형에서 반응곡선에 대한 모형은 다음과 같이 주어지며

$$\text{logit}(p(x)) = \log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x, \tag{2.2}$$

따라서 다음과 같이 표현된다.

$$p(x) = \frac{\exp(\eta(x))}{1+\exp(\eta(x))}. \tag{2.3}$$

그러므로 로짓모형에서 반응곡선에 대한 추정량은 각 개별 모수에 대한 최대가능도 추정량 $\hat{\beta}_0, \hat{\beta}_1$을 구하여 식 (2.3)에 삽입하게 되면 얻게 된다.

각 개별 모수에 대한 최대가능도 추정량을 얻기 위해서는 완전한 형태의 가능도 함수가 주어져야 한다. 그러나 어떤 경우에는 완전한 형태의 가능도 함수를 알 수는 없으나 반응변수의 평균과 분산의 관계만은 규명할 수 있는 경우가 있다. 예를 들어 반응변수의 평균과 분산이 임의의 함수 V에 의하여 $\text{Var}(Y|X=x) = V(p(x))$와 같이 표현된다고 하자. 이러한 경우에는 다음의 조건을 만족시키는 준가능도 함수를 구성하고

$$\frac{\partial}{\partial\omega} Q(\omega, y) = \frac{y-\omega}{V(\omega)}, \tag{2.4}$$

다음에 주어진 준가능도를 최대화 시킴으로 해서 모수 $\beta = (\beta_0, \beta_1)^T$를 추정할 수 있다.

$$\sum_{i=1}^{n} Q\left[g^{-1}(\beta_0 + \beta_1 X_i), Y_i\right]. \tag{2.5}$$

우리가 다루고 있는 이항반응변수의 경우에 함수 V가 $V(p) = p(1-p)$를 만족하며 이 경우에는 준가능도에 의한 추정방법이 일반적인 로그 가능도에 의한 추정방법과 동일하게 된다.

Fan 등(1995)는 준가능도 추정량을 국소화 시키는 국소선형 준가능도를 제안하였는데 그들이 제시한 목적함수는 다음과 같다.

$$l(\beta) \equiv \sum_{i=1}^{n} Q\left[g^{-1}\left(\beta_0 + \beta_1(x - X_i)\right), Y_i\right] K\left(\frac{x - X_i}{h}\right). \tag{2.6}$$

단, h는 띠폭이고 K는 커널함수이다. 목적함수 식(2.6)을 최대화시키는 $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)^T$을 구하게 되면 $\eta(x)$에 대한 국소선형 준최대가능도 추정량을 다음과 같이 구할 수 있게 되며

$$\hat{\eta}(x; d, h) = \hat{\beta}_0, \tag{2.7}$$

이어서 연결함수의 역함수를 적용시켜서 반응곡선에 대한 국소선형 준가능도 추정량을 구하게 된다.

$$\hat{p}(x; d, h) = g^{-1}\left(\hat{\eta}(x; d, h)\right). \tag{2.8}$$

2.2. 자료의 희박성 문제

자료의 희박성 문제가 추정량에 미치는 영향을 명확하게 파악하기 위해서는 추정량의 형태가 명시적으로 표현되어야 한다. 예를 들어, 자료의 희박성 문제가 국소선형 추정량에 미치는 영향은 다음과 같이 명시적으로 표현되는 국소선형 추정량의 형태에서 찾아볼 수 있다.

$$\left(\sum_{i=1}^{n} w_i Y_i\right) \Big/ \sum_{i=1}^{n} w_i. \tag{2.9}$$

단,

$$w_i(x) \equiv K\left(\frac{x-X_i}{h}\right)\{s_2-(x-X_i)s_1\}, \tag{2.10}$$

$$s_k = \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)(x-X_i)^k, \quad k=1,2, \tag{2.11}$$

커널함수 K는 $[-1,1]$에서 정의된 함수라고 하자. 이러한 경우에 만일 $[x-h, x+h]$에 X_i가 두 개 이상 존재하지 않으면 국소선형 추정량의 분모, $\sum_{i=1}^{n} w_i$의 값이 0이 되어 추정값이 정의되지 않는다.

이와 같이 자료의 희박성 문제를 명확하게 파악하기 위해서는 추정량의 형태가 명시적으로 나타나야 하는데, 식 (2.6)에 주어진 국소선형 준가능도를 최대화시키는 $\hat{\beta}=(\hat{\beta}_0,\hat{\beta}_1)^T$을 구하는 작업은 일반적으로 Fisher의 점수화방법과 같은 반복적인 계산방법을 이용해야 하기 때문에 결과적으로 얻어지는 국소선형 준가능도 추정량은 명시적 형태로 나타나지 않는다. 따라서 자료가 희박한 영역에서 국소선형 준가능도 추정량에 어떤 문제가 발생하는지를 직접적으로 파악하는 것은 쉽지 않은 문제가 된다. 그러나 국소선형 준가능도 추정량과 거의 동일한 특성을 갖고 있으면서도 명시적인 형태를 취하고 있는 추정량이 있다면 그 추정량을 통하여 국소선형 준가능도 추정량이 자료가 희박한 영역에서 문제를 갖게 되는지 여부를 간접적으로 파악할 수 있을 것이다.

국소선형 준가능도 추정량과 거의 동일한 특성을 갖고 있으면서도 명시적인 형태를 취하고 있는 추정량으로는 Fan과 Chen(1999)이 제안한 일단계 국소선형 준가능도 추정량이 있다. 식 (2.6)에 주어진 함수 $l(\beta)$의 경사도 벡터와 헤시안 행렬을 각각 $l'(\beta)$와 $l''(\beta)$라고 하자. Fan과 Chen이 제안한 일단계 국소선형 준가능도 추정량 $\tilde{\beta}=(\tilde{\beta}_0,\tilde{\beta}_1)^T$은 다음과 같이 정의된다.

$$\tilde{\beta} = \hat{\beta}^* - [l''(\hat{\beta}^*)]^{-1} l'(\hat{\beta}^*). \tag{2.12}$$

단, $\hat{\beta}^*=(\hat{\beta}_0^*,\hat{\beta}_1^*)$는 초기 추정량 벡터이며, 만일 초기 추정량 벡터 $\hat{\beta}^*$를 다음의 목적함수를 최대화시켜 구하게 되는 국소선형 회귀모형에 의한 추정량으로 사용하게 되면

$$\sum_{i=1}^{n}\{Y_i-\beta_0^*-\beta_1^*(x-X_i)\}^2 K\left(\frac{x-X_i}{h}\right), \tag{2.13}$$

식 (2.12)에 주어진 $\tilde{\beta}_0$과 식 (2.6)에서 구해지는 $\hat{\beta}_0$은 동일한 점근적 성질을 갖고 있을 뿐만이 아니라 유한표본에서도 거의 유사한 특성을 보이고 있음이 증명되었다. 그런데 식

(2.13)에서 구해지는 $\hat{\beta}_0^*$은 바로 식 (2.9)에 주어진 국소선형 추정량임을 알 수 있으며, 따라서 자료의 희박성 문제를 갖고 있는 $\hat{\beta}_0^*$과 식 (2.12)의 관계를 갖고 있는 $\tilde{\beta}_0$도 자료가 희박한 영역에서는 추정결과에 심각한 문제를 가지게 될 것이다.

자료의 희박성 문제를 확인하는 다른 방법으로 국소선형 준가능도 추정량의 조건부 및 무조건부 분산의 하한과 상한을 유도하여 분산이 무한정 큰 값을 가질 수 있음을 증명하는 방법이 있겠으나, 아직 명확하게 증명을 하지 못하였고 따라서 이 논문에서는 비록 간접적인 방법이기는 하지만 일단계 국소선형 준가능도 추정량을 통하여 자료의 희박성 문제가 발생할 수 있음을 확인하는 것으로 만족하고자 한다.

3. 자료의 희박성 문제 해결방안

자료의 희박성 문제를 해결하는 방안으로 이 논문에서는 자료가 희박한 구간에 유사자료를 적절하게 배치하는 Hall과 Turlach(1997)의 방법을 이용하고자 한다. 유사자료 (X^*, Y^*)를 적절하게 배치한다는 것은 곧 X^*와 Y^*의 값을 적절하게 결정하는 것을 의미하는 것인데 X^*의 값을 결정하는 방법은 Hall과 Turlach가 개발한 방법을 그대로 적용하는 것이 가장 최선이라고 생각되나, Y^*의 값을 결정하는 방법은 조금 더 다양한 방법을 사용할 수 있을 것으로 생각된다. 즉, Hall과 Turlach는 연속형 반응변수에 대한 국소선형 추정량의 경우만을 다루고 있기 때문에 만일 $X^* \in (X_i, X_{i+1})$이라고 한다면 Y^*의 값은 (X_i, Y_i)와 (X_{i+1}, Y_{i+1})의 선형보간법으로 결정하는 것이 가장 적합한 방법이 된다고 하겠다. 그러나 이항반응변수의 경우에는 이웃한 두 관찰값의 선형보간법 이외에 좀더 다양한 방법을 시도해 볼 수 있을 것으로 생각된다.

유사자료의 개수 및 X^*값을 결정하는 방법으로 Hall과 Turlach가 제안한 방법은 다음과 같다. 설명변수 X_i는 밀도함수 f를 갖고 있는 확률변수인데 $a \le X_1 \le \cdots \le X_n \le b$의 관계를 만족하고 있으며, 커널함수 K는 $[-1, 1]$에서만 정의되는 함수라고 가정하자. 또한 $X_0 = a$, $X_{n+1} = b$라고 하면 $S_i = X_{i+1} - X_i$, $0 \le i \le n$은 자료의 i번째 구간을 나타내는 것이 되며, $\mathcal{I}_h$를 $a + h \le X_i \le X_{i+1} \le b - h$를 만족하는 첨자 i의 집합이라고 하자. 주어진 상수 r에 대한 m_i의 값을 만일 $i \in \mathcal{I}_h$인 경우에는 $rS_i/2h$의 정수 부분의 값이라고 하고, 그 이외의 경우에는 rS_i/h의 정수 부분의 값이라고 하자. 유사자료를 배치하게 될 위치는 m_i의 값에 의하여 결정이 되는데 m_i의 값이 1 이상이 되는 경우에만 (X_i, X_{i+1})의 구간에 m_i개의 유사자료를 균등간격으로 배치하게 된다. 이 방법으로 유사자료를 배치하게 되면 $[a, b]$ 사이에 존재하는 모든 x에 대하여 $(x - h, x + h)$의 구간에는 적어도 상수 r의 정수 부분의 값만큼의 자료 또는 유사자료가 존재하게 되어 자료가 희박한 영역이 없어지게 된다.

이제 유사자료 Y^*값을 결정하는 방법에 대해서 생각해 보자. 우선 $Y_0 \equiv Y_1$이고 $Y_{n+1} \equiv Y_n$이라 하고 유사자료 X^*값은 $X^* \in (X_i, X_{i+1})$에 있는 것으로 결정되었다고 하자. 실제자료 Y는 0 또는 1의 값만을 갖고 있지만 유사자료 Y^*는 (0,1) 구간에 속하는 값을 가질 수 있도록 허용하여 주변에 있는 실제자료들의 특성을 더 잘 반영할 수 있도록 융통성을 부여하는 것이 더 효과적이라고 생각된다. 주변에 있는 실제자료를 이용하는 방법에 따라서 다르게 정의되는 다음의 다섯 가지 방법을 제안하고자 한다.

M_1: $Y^* = 0.5$.

M_2: $Y^* = (Y_i + Y_{i+1})/2$.

M_3: (X_i, Y_i)와 (X_{i+1}, Y_{i+1})의 선형보간법으로 Y^*의 값 결정.

M_4: 최근접이웃방법에 의한 국소선형 준가능도 추정량으로 반응곡선 추정값 $\hat{p}(x)$을 우선 구하고, 이어서 $Y^* = \hat{p}(X^*)$으로 결정

M_5: 고정 띠폭에 의한 국소선형 준가능도 추정량으로 반응곡선 추정값 $\hat{p}(x)$을 우선 구하고, 이어서 $Y^* = \hat{p}(X^*)$으로 결정

방법 M_1은 이웃한 자료의 정보를 전혀 이용하지 않고 모든 유사자료 Y^*값을 0.5로 고정하는 방법이다. 가장 간단한 방법이지만 주변에 있는 실제자료들의 특성을 제대로 살리지 못할 가능성도 있을 것이다. 방법 M_2와 M_3는 이웃한 두 실제자료만을 이용하는 방법이다. 만일 추가되는 유사자료의 수가 한 개일 경우에는 동일한 결과를 낳게 되지만 2개 이상의 유사자료가 (X_i, X_{i+1})의 구간에 배치된다면 다른 결과가 나올 수도 있게 된다. 방법 M_4와 M_5는 국소선형 준가능도 추정량으로 1차 추정된 반응곡선을 이용하는 방법이다. 주변자료의 정보를 가장 폭 넓게 이용하고 있으나 1차 추정된 반응곡선이 자료가 희박한 부분에서는 왜곡된 결과를 보일 수 있기 때문에 그 나름대로의 한계가 있는 방법이라고 하겠다.

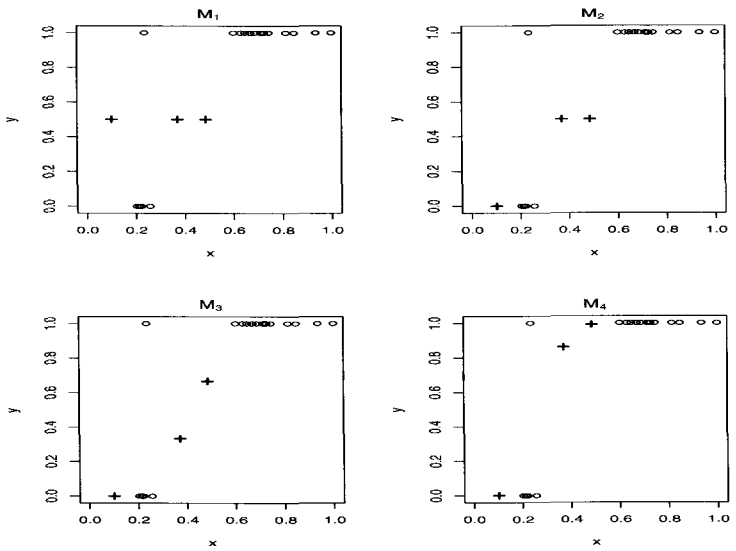

그림 3.1: 유사자료 Y^*의 값을 결정하는 네 가지 방법을 전형적인 자료에 적용시켜 얻은 결과의 비교.(실제자료는 작은 원으로 표시되어 있고 추가된 유사자료는 십자로 표시되어 있다.)

그림 3.1은 M_1에서 M_4까지 네 가지 방법을 전형적인 자료에 적용시켜 얻은 결과를 나타낸 것이다. 자료는 $n = 20$으로 Uniform(0, 1)에서 임의로 X변수의 값을 추출하였고 Y변수는 반응곡선 $p(x) = [1 + \exp(5 - 15x)]^{-1}$에서 얻었다. 상수 $r = 3$이 사용되었고 띠폭으로는 0.4가 사용되었다. 2개의 구간에 모두 3개의 유사자료가 각 방법에 따라서 위치가 결정되었다. M_5의 경우는 M_4와 거의 비슷한 결과를 보였기 때문에 생략하였다. 각 방법의 특징이 나름대로 잘 나타나 있다.

4. 모의실험

고정된 띠폭을 사용하는 국소선형 준가능도 추정량은 경우에 따라서 자료의 희박성 문제를 겪을 수 있다는 것이 지적되었고, 이 문제를 해결하는 방법으로 유사자료를 적절하게 배치하는 다섯 가지 방법이 앞 절에서 제안되었다. 이 방법들 중에 어떤 방법이 가장 효율적인지에 대한 답을 이론적으로 도출하는 것은 매우 어려운 작업이 될 것으로 생각된다. 더욱이 자료의 희박성 문제는 자료의 크기가 상대적으로 작을 때 빈번하게 발생할 것으로 예상이 되기 때문에 점근적인 비교는 큰 의미를 갖지 못할 것이다. 따라서 가장 효과적인 비교방법은 모의실험을 통한 것이라고 하겠다.

모의실험은 다음 여섯 가지의 모형을 실제반응곡선으로 사용하여 실시하였다.

1. 로짓모형, $p_1(x) = [1 + \exp(5 - 15x)]^{-1}$
2. 보 로그-로그모형, $p_2(x) = 1 - \exp(-\exp(-5 + 8x))$
3. 정규혼합모형, $p_3(x) = 0.5\Phi(\frac{x-0.3}{0.05}) + 0.5\Phi(\frac{x-0.7}{0.05})$
4. 와이블모형, $p_4(x) = 1 - \exp(-(15x)^{0.5})$
5. 삼차로지스틱모형, $p_5(x) = [1 + \exp(-(2(x - 0.5) - 40(x - 0.5)^3))]^{-1}$
6. 이단계모형, $p_6(x) = 1 - \exp[-0.5 - 3.15(x - 0.5) - 5(x - 0.5)^2]$

그림 4.1에서 볼 수 있듯이 모형 $p_1(x)$부터 $p_4(x)$까지의 반응곡선은 비감소함수의 형태를 취하고 있는데, 이와 같이 비감소함수의 형태를 지니고 있는 반응곡선에 대해서는 Müller와 Schmitt(1988)에서 언급된 것처럼 반응곡선 $p(x)$의 추정뿐만이 아니라 $p^{-1}(\alpha), 0 < \alpha < 1$의 추정도 중요한 연구목적이 된다. 따라서 반응곡선의 추정값도 반드시 비감소함수의 조건을 만족시키는 것이 $p^{-1}(\alpha)$의 추정을 위하여 필요하다고 하겠다. 그러나 국소선형 준가능도 추정값 $\hat{p}(x)$는 비감소 조건을 만족시킨다는 보장을 할 수 없으며, 따라서 이러한 경우에 적절한 대안이 필요하게 된다. 즉, $x_1 \le x_2 \le \cdots \le x_n$에서 계산된 국소선형 준가능도 추정값 $\hat{p}(x_i)$, $i = 1, \ldots, n$에 대하여 $\tilde{p}(x_1) \le \tilde{p}(x_2) \le \cdots \le \tilde{p}(x_n)$을 만족하면서 $\sum_{i=1}^{n} (\hat{p}(x_i) - \tilde{p}(x_i))^2$을 최소화시키는 $\tilde{p}(x_i)$를 찾는다면 $\hat{p}(x_i)$ 대신에 $\tilde{p}(x_i)$를 우리의 추정값으로 대신 사용할 수 있을 것이다. 추정값 $\tilde{p}(x_i)$는 PAVA(Pool Adjacent Violators Algorithm)로 얻을 수 있다.

표본크기는 $n = 10, 20, 30, 40, 50, 100$이 사용되었으며, 설명변수의 설계점 X_i들은 Uniform(0,1)의 분포에서 난수를 발생하여 결정하였고, 각 모형에 대한 반응변수의 값 Y_i는 Uniform(0,1)에서 발생시킨 난수와 $p(X_i)$를 비교하여 결정하였다.

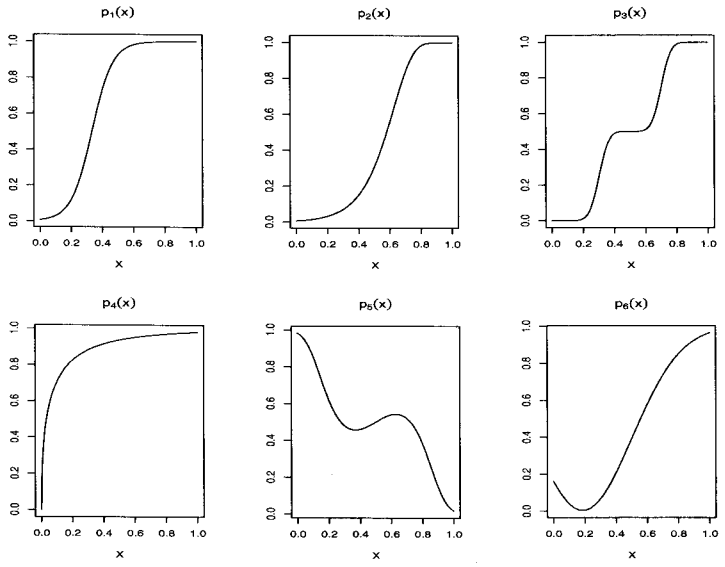

그림 4.1: 모의실험에서 사용된 반응곡선의 모습

표본이 주어지면 우선 고정된 띠폭과 최근접이웃 방법에 의하여 각각 반응곡선을 추정하였고, 이어서 다섯 가지 방법 각각에 의하여 유사자료를 추가시킨 자료를 마련하였다. 유사자료의 X값을 결정하는데 중요한 요소인 r의 값은 Hall과 Turlach(1997)에서 추천한 $r = 3$를 사용하였다. 이어서 다섯 가지 각 방법에 의하여 생성된 자료에 대하여 고정된 띠폭으로 반응곡선을 추정하였다. 따라서 모두 7개의 반응곡선의 추정값이 비교대상이 되는 것이다.

국소선형 준가능도 추정량에 의한 반응곡선의 추정값, $\hat{p}(x)$는 R함수 *locfit*을 사용하여 계산하였고, 커널함수는 Epanechnikov 함수를 사용하였다.

각 방법의 효율성 비교를 위한 ISE(Integrated Squared Error)를 계산하기 위하여 우선 각 방법에 의한 자료에 대하여 [0.1, 0.9] 구간의 21개 격자점으로 구성된 띠폭을 하나씩 이용하여 반응곡선을 추정하였다. 모형 $p_1(x)$부터 $p_4(x)$에서는 만일 추정된 반응곡선이 비감소의 조건을 만족시키지 않는 경우에 PAVA로 변환을 시켰다. 이어서 [0, 1]의 구간에 대한 100개의 격자점에서 예측된 반응곡선의 추정값을 대상으로 ISE를 계산하였다.

따라서 각 방법마다 21개의 ISE가 계산되는데, 이 중에 가장 작은 값을 주어진 표본에 대한 각 방법의 ISE로 할당하였다. 즉, ISE를 최소화시키는 띠폭을 선택하여 반응곡선을 추정한 것이 된다. 실제 상황에서는 사용할 수 없는 방법으로 띠폭을 선택한 이유는 cross-validation이나 plug-in과 같은 자료적응방법에 의하여 선택된 띠폭이 최적이라는 보장이 없기 때문이다. 잘못 선택된 띠폭 때문에 왜곡된 비교결과가 나올 가능성을 없애기 위한 방안이라고 하겠다.

이어서 각 방법에 대한 효율성 비교는 MISE로 실시하였는데, MISE는 1000번의 반복으

로 얻어진 ISE의 평균값으로 계산하였다. 각 방법에 대한 MISE는 표 4.1에 있다.

모의실험 결과에서 우선 눈에 띄는 것은 최근접이웃 방법의 MISE 값이 고정된 띠폭에 의한 방법과 거의 동일하거나 혹은 더 큰 값을 갖고 있다는 것이다. 이것은 최근접이웃 방법이 이항반응변수에 대한 국소선형 준가능도 추정량이 갖고 있는 자료의 희박성 문제를 해결하는 대안이 전혀 될 수 없다는 것을 의미하는 것이 된다.

제안된 다섯 가지 방법의 효율성을 비교해 보자. 우선 비감소 조건을 만족시키기 위하여 PAVA에 의한 단조변환이 필요했던 모형 $p_1(x)$부터 $p_4(x)$의 경우를 살펴보면 $n = 10$에서는 이웃한 두 개의 실제자료를 이용하는 M_2와 M_3가 다른 방법보다 더 좋은 결과를 보였으나 $n = 20, 30, 40$에서는 가장 단순한 방법인 M_1이 가장 좋은 결과를 보이고 있다. 또한 단조변환이 필요하지 않은 모형 $p_5(x)$와 $p_6(x)$의 경우에는 $n = 10, 20, 30, 40$에서 모두 M_1이 가장 좋은 결과를 보이고 있다. 특히 $n = 10$에서는 상대적으로 매우 탁월한 결과를 보이고 있다.

이웃한 자료의 정보를 전혀 이용하지 않고 있는 가장 단순한 방법인 M_1이 많은 경우에 있어서 상당히 좋은 결과를 보이고 있는 점은 매우 의외라고 하겠으며, 주변자료의 정보를 가장 포괄적으로 이용하고 있는 M_4와 M_5의 효과가 거의 없는 것으로 나타난 것도 예상 밖의 결과라고 하겠다. 그러나 M_4와 M_5의 경우 유사자료 Y^*의 값을 결정하는데 사용된 반응곡선의 추정값이 예상보다 많이 왜곡되어 있음을 표 4.1를 통하여 알 수 있었고 따라서 두 방법이 좋지 못한 결과를 보이는 것은 어찌보면 당연하다고 하겠다.

또한 이웃한 자료의 정보를 효과적으로 이용하는 방법인 M_2와 M_3가 기대보다 못한 결과를 보인 것은 이항반응변수의 특성때문이라고 하겠다. 연속형 반응변수의 경우에는 이웃한 자료끼리 많은 정보를 서로 공유하고 있는 것이 사실이겠으나, 0과 1의 값만을 갖고 있는 이항반응변수의 경우에는 정보의 공유 정도가 상대적으로 약하다고 할 수 있겠다. 따라서 이항반응변수의 경우에 주어진 자료의 정보를 이용함으로써 얻는 효과는 연속형 반응변수의 경우만큼 크지 않다고 할 수 있겠다. 더욱이 Y_i와 Y_{i+1}이 이상값인 경우에 M_2와 M_3에 의하여 생성된 유사자료 역시 이상값이 되기 때문에 결국 한 무리의 이상값을 자료에 추가한 꼴이 되어 결과적으로 매우 큰 ISE 값을 갖게 됨을 모의실험 과정에서 확인할 수 있었다. 반면에 M_1은 주변의 이상값과는 관계없이 항상 0.5의 값을 부여함으로 해서 최악의 결과를 방지할 수 있었다. 이렇듯 주어진 자료의 정보를 전혀 이용하지 않는 가장 단순한 방법인 M_1이 비록 항상 최고의 결과를 주는 최선의 방법은 아닐지라도 최대 손실을 최소화시킬 수 있는 방법이기 때문에 가장 작은 MISE의 값을 갖게 된것이라고 보인다.

또한 $n = 50$의 경우부터는 방법들 간에 큰 차이가 없어지기 시작했으며, $n = 100$의 경우에는 모든 방법이 실질적으로 동일한 결과를 보이고 있다. 이것은 자료의 희박성 문제가 표본크기가 작은 경우에만 중요한 논점이 된다는 것을 의미하는 것으로 보인다.

5. 결론

국소선형 추정량은 여러 면에서 바람직한 특성을 많이 갖고 있는 좋은 추정량이다. 그러나 자료가 희박한 부분에서는 매우 불안정한 추정값을 갖게 되는 문제를 지니고 있음이

표 4.1: 일곱 가지 추정방법에 대한 MISE

p	n	M_1	M_2	M_3	M_4	M_5	최근접	고정
	10	.0181	.0164	.0177	.0176	.0188	.0189	.0190
	20	.0083	.0099	.0090	.0107	.0111	.0110	.0111
	30	.0065	.0077	.0070	.0080	.0080	.0082	.0081
p_1	40	.0057	.0064	.0063	.0065	.0066	.0067	.0066
	50	.0048	.0052	.0052	.0053	.0053	.0054	.0053
	100	.0024	.0024	.0024	.0024	.0024	.0025	.0024
	10	.0180	.0189	.0163	.0202	.0215	.0220	.0217
	20	.0096	.0119	.0105	.0124	.0127	.0133	.0127
	30	.0074	.0088	.0079	.0090	.0090	.0095	.0091
p_2	40	.0068	.0073	.0069	.0073	.0074	.0077	.0074
	50	.0054	.0055	.0054	.0056	.0056	.0059	.0056
	100	.0029	.0029	.0029	.0029	.0029	.0030	.0029
	10	.0202	.0236	.0211	.0248	.0256	.0271	.0262
	20	.0150	.0180	.0160	.0185	.0186	.0190	.0187
	30	.0133	.0147	.0137	.0149	.0150	.0152	.0150
p_3	40	.0111	.0117	.0111	.0117	.0118	.0119	.0118
	50	.0098	.0100	.0097	.0100	.0100	.0100	.0100
	100	.0060	.0060	.0060	.0060	.0060	.0059	.0060
	10	.0384	.0373	.0402	.0395	.0409	.0410	.0413
	20	.0142	.0163	.0159	.0169	.0173	.0176	.0174
	30	.0099	.0105	.0103	.0108	.0109	.0112	.0109
p_4	40	.0080	.0083	.0082	.0085	.0085	.0088	.0085
	50	.0069	.0070	.0069	.0070	.0070	.0074	.0070
	100	.0045	.0045	.0045	.0045	.0045	.0048	.0045
	10	.1033	.1731	.1683	.1691	.1683	.1899	.1690
	20	.0630	.0761	.0748	.0750	.0750	.0830	.0752
	30	.0421	.0450	.0447	.0446	.0445	.0488	.0447
p_5	40	.0285	.0290	.0290	.0290	.0290	.0313	.0291
	50	.0222	.0223	.0223	.0223	.0223	.0241	.0223
	100	.0103	.0103	.0103	.0103	.0103	.0108	.0103
	10	.0180	.0219	.0196	.0236	.0248	.0278	.0251
	20	.0123	.0149	.0134	.0154	.0159	.0177	.0160
	30	.0101	.0111	.0103	.0113	.0113	.0125	.0113
p_6	40	.0082	.0086	.0082	.0086	.0087	.0095	.0087
	50	.0068	.0068	.0068	.0069	.0069	.0075	.0069
	100	.0038	.0038	.0038	.0038	.0038	.0041	.0038

밝혀졌으며, 이 문제를 해결하기 위한 여러 방안이 많이 연구되었다. 그러나 이항반응변수를 위한 국소선형 추정량의 변형이라고 할 수 있는 국소선형 준가능도 추정량에 대해서는 아직 자료의 희박성 문제가 다루어지지 않고 있었다. 이 논문에서는 국소선형 준가능도 추정량이 갖고 있는 자료의 희박성 문제를 인식하고, 유사자료의 위치선정에 관련하여 5가지 방안을 제시하였으며, 모의실험을 통하여 제시된 방안들의 효율성을 비교하였다.

모의실험 결과 가장 효과적인 방법으로 채택된 M_1은 이웃한 자료의 정보를 전혀 이용하지 않는 방법이어서 가장 효과가 없을 방법이라고 예상을 하였으나, 결과는 정반대로 나왔다. 그러나 이러한 결과는 반응변수가 이항자료이기 때문에 실질적으로 이웃한 자료끼리 많은 정보를 서로 공유하기 어렵다는 점을 감안한다면 큰 무리가 없는 결과라고 본다.

또한 자료의 희박성 문제는 표본크기가 상당히 큰 경우에도 여전히 중요한 논점이 될 수 있을 것이라는 Seifert와 Gasser(1996)의 지적이 이항반응변수의 경우에는 적용이 되지 않는다는 것도 새롭게 밝혀진 점이라고 하겠다.

이와 같은 발견은 이항자료의 희박성 문제가 발생했을 때 연속형 자료의 경우에 대한 접근방법으로는 해결이 어렵다는 것을 의미하는 것이며 따라서 이항자료만을 위한 접근방법의 존재가치를 드러내는 것이라고 하겠다.

참고문헌

Fan, J. and Chen, J. (1999). One-step local quasi-likelihood estimation, *Journal of the Royal Statistical Society*, Ser. B, **61**, 927–943.

Fan, J., Heckman, N. E. and Wand, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi–likelihood functions, *Journal of the American Statistical Association*, **90**, 141–150.

Hall, P. and Turlach, B. A. (1997). Interpolation method for adapting to sparse design in nonparametric regression, *Journal of the American Statistical Association*, **92**, 466–477.

Jennen-Steinmetz, C. and Gasser, T. (1988). A unifying approach to nonparametric regression estimation, *Journal of the American Statistical Association*, **83**, 1084–1089.

Müller, H. and Schmitt, T. (1988). Kernel and probit estimates in quantal bioassay, *Journal of the American Statistical Association*, **83**, 750–759.

Park, D. (1999). Comparison of two response curve estimators, *Journal of Statistical Computation and Simulation*, **62**, 259–269.

Park, D. and Park, S. (2006). Parametric and nonparametric estimators of ED100α, *Journal of Statistical Computation and Simulation*, **76**, 661–672.

Seifert, B. and Gasser, T. (1996). Finite-sample variance of local polynomials: Analysis and solutions, *Journal of the American Statistical Association*, **91**, 267–275.

[2006년 8월 접수, 2006년 11월 채택]

Sparse Design Problem in Local Linear Quasi-likelihood Estimator*

Dongryeon Park[1)]

ABSTRACT

Local linear estimator has a number of advantages over the traditional kernel estimators. The better performance near boundaries is one of them. However, local linear estimator can produce erratic result in sparse regions in the realization of the design and to solve this problem much research has been done. Local linear quasi-likelihood estimator has many common properties with local linear estimator, and it turns out that sparse design can also lead local linear quasi-likelihood estimator to erratic behavior in practice. Several methods to solve this problem are proposed and their finite sample properties are compared by the simulation study.

Keywords: Binary response variable, local linear quasi-likelihood estimator, Pseudo data.

* This research was supported by Hanshin University Research Grant in 2007.

1) Professor, Dept. of Statistics, Hanshin University, Osan, Kyunggi-do 447-791, Korea
E-mail: drpark@hs.ac.kr