응용통계연구 제 **20**권 *3*호, *2007*년, *pp. 561–571*

선형보간법에 의한 자료 희소성 해결방안의 문제와 대안*

박동련[1)]

요 약

국소선형회귀모형의 추정량은 좋은 특성을 가지고 있는 추정량으로서 가장 흔히 사용되는 비모수적 회귀모형의 추정량이라고 하겠다. 이러한 국소선형 추정량이 자료가 희박한 구간에서는 심하게 왜곡된 추정결과를 보이는 문제가 있으며, Hall과 Turlach (1997)이 제안한 선형보간법이 이러한 문제에 대한 매우 효과적인 해결방안이라는 것은 잘 알려진 사실이다. 그러나 Hall과 Turlach가 제안한 선형보간법이 이상값에 매우 취약하다는 사실은 아직 지적된 적이 없는 문제이다. 이 논문에서는 이상값의 영향력을 감소시킬 수 있는 수정된 선형보간법에 의한 유사자료의 생성방법을 제안하고, 그 특성을 모의실험을 통하여 기존의 방법과 비교하였다.

주요용어: 국소선형 회귀모형, 로버스트성 가중값, 선형보간법, 유사자료, 이상값.

1. 서론

주어진 자료 $\{(X_i, Y_i)\}_{i=1}^n$를 이용하여 회귀함수를 추정하는 문제를 생각해보자. 두 변수 (X, Y) 사이에 다음의 관계가 있다고 가정하고

$$Y_i = g(X_i) + \epsilon_i, \quad i = 1, \ldots, n \tag{1.1}$$

회귀함수 g의 형태에 대해서는 매끄럽게 (smooth) 생겼다는 것 이외에는 어떠한 가정도 하지 않는다고 하자. 이러한 경우에 회귀함수 g을 추정하는 문제를 비모수적 회귀문제라고 하는데 매우 다양한 방법들이 제안되어 있다. 많이 사용되는 방법으로는 커널 추정량, 직교열 추정량, 평활 스플라인 추정량 그리고 국소다항 추정량 등이 있다. 이 방법들은 모두 각기 장점들을 가지고 있는데, 일반적으로 국소다항 추정량이 가장 좋은 특성을 갖고 있는 추정량으로 알려져 있다. 국소다항 추정량의 특성에 대한 자세한 설명은 Fan (1992, 1993), Fan과 Gijbels (1996), 그리고 Wand와 Jones (1995) 등에서 찾아 볼 수 있다.

국소다항 추정량에서 차수가 일차인 국소선형 추정량이 일반적으로 많이 사용되고 있는데, 국소선형 추정량은 비록 좋은 특성을 많이 가지고 있는 추정량이기는 하지만 나름대로의 문제점도 가지고 있다. Seifert와 Gasser (1996)는 국소선형 추정량을 실제 자료에 적용시켰을 때 기대했던 것보다 훨씬 나쁜 추정결과가 종종 나오게 되는 현상을 지적했는데, 그들은 이 문제가 자료의 희소성 (sparsity) 때문에 발생하는 것임을 밝혔다.

* 이 논문은 2007년도 한신대학교 학술연구비 지원에 의하여 연구되었음.

1) (447-791) 경기도 오산시 양산동 411, 한신대학교 정보통계학과, 교수
E-mail: drpark@hs.ac.kr

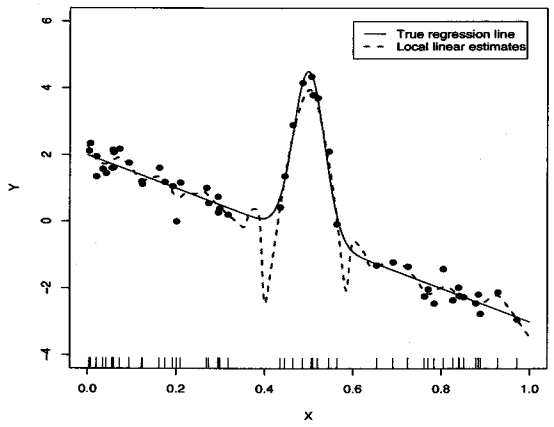

그림 1.1: 국소선형 추정량이 자료가 희박한 구간에서 비정상적인 추정 결과를 보이는 현상

그림 1.1은 자료가 희박한 구간에서 국소선형 추정량이 매우 심하게 왜곡된 추정결과를 보일 수 있음을 예시하고 있다. X 관찰값은 Uniform$(0,1)$의 분포에서 임의로 $n=50$의 자료를 추출하여 얻었고 Y 관찰값은 다음의 회귀함수 $g(x)$에 $N(0,0.5)$의 오차를 더하여 얻었다.

$$g(x) = 2 - 5x + 5\exp\left[-400(x-0.5)^2\right], \quad x \in [0,1]. \tag{1.2}$$

$x=0.4$와 $x=0.6$ 근처에서 자료가 희박한 구간이 있음을 알 수 있으며, 그 구간에서는 추정된 회귀선이 많이 왜곡되어 있음도 알 수 있다.

이러한 자료의 희소성 문제를 해결하는 방법으로 Seifert와 Gasser는 국소선형 추정량을 정의하는 과정에서 능형모수를 포함시키는 방법을 제시하였다. 그러나 이 방법은 상대적으로 복잡할 뿐더러 능형모수를 선택해야 하는 부가적인 문제를 안고 있게 된다. Hall과 Turlach (1997)는 자료가 희박한 구간에 선형보간법으로 생성된 유사자료를 원자료에 포함시킴으로 해서 이 문제가 간단하게 해결될 수 있음을 보였고, 모의실험을 통하여 그들의 방법이 Seifert와 Gasser (1996)의 능형모수를 이용하는 방법보다 더 효율적임을 보였다.

그림 1.2는 그림 1.1의 원자료에 Hall과 Turlach가 제안한 방법으로 생성된 유사자료를 포함시켜 국소선형 추정량으로 회귀곡선을 추정한 결과를 보여주고 있다. 자료의 희소성 문제가 완벽하게 해결되었음을 알 수 있다. 그림 1.1과 그림 1.2 모두에서 사용된 커널함수는 Epanechnikov 함수이며 띠폭은 MISE를 최소화시키도록 설정된 $h=0.0457$를 사용하였다.

Hall과 Turlach가 제안한 방법을 간단하게 살펴보면 다음과 같다. 만일 $a \le X_1 \le \cdots \le X_n \le b$이라고 하였을 때 우선 (X_i, X_{i+1})의 구간에 유사자료가 필요한지 여부를 결정하게 되는데, 이것은 모든 $x \in [a,b]$에 대하여 $(x-h, x+h)$의 구간에 적어도 r개의 자료가 포함되어 있는지 여부를 확인하는 단계가 된다. 만일 어떤 x에 대하여 $(x-h, x+h)$의 구간에

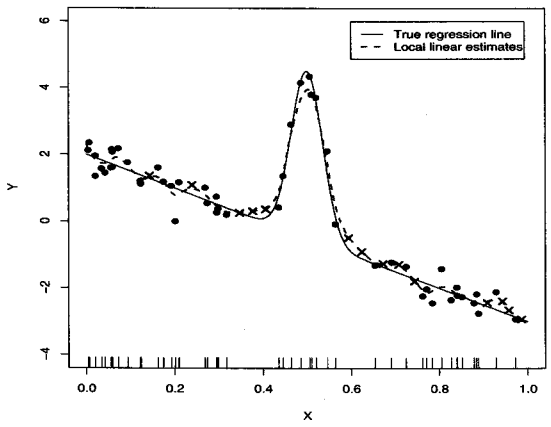

그림 1.2: 선형보간법으로 생성된 유사자료를 포함시킴으로써 자료의 희소성 문제가 해결될 수 있음을 보이고 있음 (추가된 유사자료는 십자표시로 나타나 있음.)

r개의 자료가 존재하지 않는다면 x가 포함된 (X_i, X_{i+1})의 구간에 필요한 만큼의 유사자료를 생성시켜야 한다. 이러한 절차에 의하여 각 (X_i, X_{i+1})의 구간마다 필요한 유사자료의 개수가 결정되면, 이어서 유사자료 (X^*, Y^*)의 위치를 결정하게 되는데, 우선 X^*의 값은 (X_i, X_{i+1})의 구간을 필요한 유사자료의 개수만큼 동일한 간격으로 나누어 계산되고, Y^*의 값은 (X_i, Y_i)와 (X_{i+1}, Y_{i+1})의 두 점을 이용한 선형보간법으로 계산된다. 상수 r의 적절한 값으로는 $r = 3$을 추천하였다.

Hall과 Turlach가 제안한 선형보간법에 의한 방법은 매우 간단하면서도 효과적이기 때문에 말그대로 이상적인 방법이라고 하겠다. 그러나 이러한 결과는 선형보간법에 이용되는 Y_i와 Y_{i+1}이 정상적인 관찰값인 경우에만 적용된다고 하겠다. 즉, 만일 Y_i와 Y_{i+1}이 이상값이라고 한다면 두 점을 이용한 선형보간법으로 생성되는 유사자료들은 모두 이상값이 될 가능성이 높게 될 것이다. 그림 1.3은 그림 1.1에서 사용된 원자료 중 하나의 Y_i값을 이상값으로 변화시킨 후 선형보간법으로 유사자료를 생성시켜 원자료에 포함시키고 국소선형 추정법으로 회귀곡선을 추정한 결과이다. $x = 0.4$ 부근에서 매우 잘못된 추정결과를 보이고 있음을 알 수 있다. 이것은 추가된 유사자료들이 이상값에 의하여 생성되었고, 따라서 회귀함수에 대한 잘못된 정보를 주고 있기 때문이다.

우리가 실제로 접하게 되는 자료에 이상값이 포함될 가능성은 충분히 존재하는 것이고 만일 이러한 이상값이 선형보간법의 기준값이 된다면 Hall과 Turlach의 방법은 다수의 이상값을 양산하는 최악의 결과를 얻게 될 것이다. 물론 2차원 공간에서의 이상값 존재여부는 산점도를 그려보는 것만으로 쉽게 확인할 수 있는 문제다. 그러나 국소선형 회귀모형을 여러 차례 적합시켜야 하는 복잡한 형태의 자료에 대해서는 항상 모든 산점도를 그려서 확인하는 것이 가능하지 않을 수도 있는 것이다. 따라서 일종의 안전판 역할을 할 수 있는 대

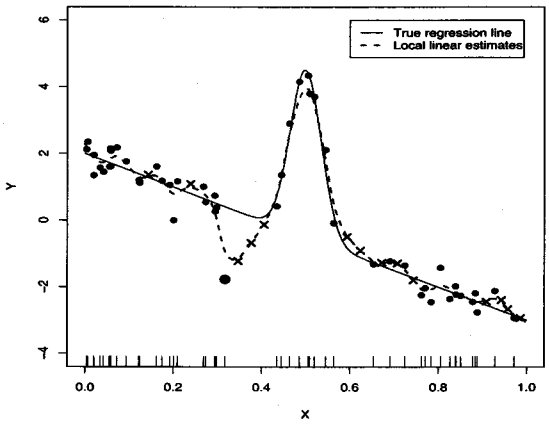

그림 1.3: $Y_{25} = 0.205$의 값을 $Y_{25} = -1.766$로 변화시킨 후 선형보간법으로 유사자료를 생성시켜 원자료에 포함시키고 국소선형 추정량으로 추정한 회귀곡선 (변화된 원자료는 큰 점으로 표시되었음.)

안이 필요하다고 하겠다. 사실 자료의 희소성 문제도 산점도를 그려보면 미리 쉽게 확인 할 수 있는 문제이지만, Hall과 Turlach가 선형보간법에 의한 방안을 제안한 것도 국소선형 추정량에 대한 안전판을 제공하기 위한 것임을 그들도 밝히고 있다. 그러나 이상값이 존재하는 경우에 그들의 방법은 더 이상 그들이 원하는 역할을 수행할 수 없기 때문에 이상값의 문제를 자동으로 해결해 줄 수 있는 또 다른 안전판이 필요하다고 하겠다.

이 논문에서는 국소선형회귀모형의 추정량이 겪게 되는 자료 희소성 문제를 해결하기 위하여 Hall과 Turlach가 제안한 선형보간법이 이상값이 존재하는 경우에는 왜곡된 결과를 낳을 수 있음을 인식하고 그 대안으로서 이상값의 영향을 감소시킬 수 있는 수정된 선형보간법을 제안하고자 한다.

2. 수정된 선형보간법

Hall과 Turlach (1997)가 제안한 선형보간법에 의한 자료 희소성 해결방안이 문제가 되는 경우는 X_i와 X_{i+1} 사이의 간격이 너무 벌어져서 유사자료를 포함시켜야 하는데 Y_i 또는 Y_{i+1}이 이상값이 되는 경우로서 이 때에는 (X_i, Y_i)와 (X_{i+1}, Y_{i+1})의 선형보간법에 의하여 생성되는 유사자료들의 다수가 이상값이 될 가능성이 높아지기 때문이다. 따라서 자료 $\{(X_i, Y_i)\}_{i=1}^n$가 주어지면 유사자료를 생성하기 전에 미리 이상값을 식별하고, 식별된 이상값들을 선형보간법에 의한 유사자료의 생성과정에 포함시키지 않는 사전 작업이 필요하다고 할 수 있으며, 그러한 작업이 선행된다면 잘못된 유사자료, 즉 이상값을 자료에 포함시키는 우를 범하지 않을 것이다.

이상값을 식별하는데 일반적으로 사용되는 방법은 적합된 모형의 잔차를 이용하는 것이다. 그러나 우리는 유사자료를 생성하기 전에 이상값을 식별해야 하는데, 이것은 곧 국소선형회귀모형을 적합시키기 전에 이상값을 탐지해야 한다는 것을 의미하는 것이며, 따라서 잔차를 이용하여 이상값을 식별하는 방법은 적용대상이 되지 못한다.

잔차를 이용하지 않고 이상값을 식별하는 방법으로 자료들의 상대적인 위치를 고려하여 판단하는 방법을 생각해 볼 수 있다. 즉, 이상값이라는 것이 주변의 자료와는 다른 성격을 지니고 있는 관찰값이기 때문에 각 관찰값들이 주변의 자료와 얼마나 잘 어울리는 지를 나타내는 측도가 있다면 이상값을 식별해 내는 것이 가능할 것이다. 이러한 목적에 잘 부합하는 방법으로 Park (2004)이 제안한 방법이 있다. 이 방법은 국소선형회귀모형을 적합시킬 때 이상값의 영향력을 감소시키기 위하여 각 관찰값 (X_i, Y_i)에 대한 로버스트성 가중값 (robustness weight) r_i를 계산하는 방법으로서, 이 가중값은 주변자료와의 상대적 위치를 고려하여 계산된다. 즉, 만일 i번째 관찰값이 주변자료와 동떨어져 있는 위치에 있다면 0에 가까운 가중값을 부여하는 것이다. 이렇게 되면 모형을 적합 시킬 때 (X_i, Y_i)는 거의 아무런 역할을 할 수 없게 되어 이상값의 영향력을 감소시킬 수 있게 된다. Park (2004)이 제안한 로버스트성 가중값 계산 절차는 다음과 같다.

1. 각 $i = 1, \ldots, n$마다 $d_j(Y_i) = |Y_i - Y_j|$와 $w_j(X_i) = T(\Delta_j(X_i)/\Delta_{(k)}(X_i))$, $j \in J_i^k$을 계산한다. 단, $J_i^k = \{j : X_j$는 X_i와 가까운 k개의 관찰값 중의 하나$\}$. 또한 함수 T는 다음과 같이 정의되며

$$T(x) = \begin{cases} \left(1 - |x|^3\right)^3, & \text{if } |x| < 1, \\ 0, & \text{otherwise}, \end{cases} \tag{2.1}$$

   $\Delta_i(x) = |X_i - x|$이고, $\Delta_{(k)}(x)$는 $\Delta_i(x)$ 중 k번째로 작은 값을 의미한다.

2. $d_j(Y_i)$, $j \in J_i^k$의 가중 중위수 m_i를 다음과 같이 구한다.

$$m_i = \begin{cases} \left[d_{(j)}(Y_i) + d_{(j+1)}(Y_i)\right]/2, & \text{if } \sum_{l=1}^{j} w_{(l)}(X_i) = 1/2, \\ d_{(j)}(Y_i), & \text{if } \sum_{l=1}^{j-1} w_{(l)}(X_i) < 1/2 < \sum_{l=1}^{j} w_{(l)}(X_i), \end{cases}$$

   단, $d_{(j)}(Y_i)$는 $d_j(Y_i)$를 오름차순으로 정리한 순서통계량이며 $w_{(j)}(X_i)$는 $d_j(Y_i)$의 크기 순에 따라서 재배열된 것을 의미한다.

3. 관찰값 (X_i, Y_i)의 로버스트성 가중값 r_i를 다음과 같이 정의한다.

$$r_i = B\left(\frac{m_i}{6s_i}\right). \tag{2.2}$$

   단, 함수 B는 다음과 같이 정의되며,

$$B(x) = \begin{cases} \left(1 - x^2\right)^2, & \text{if } |x| < 1, \\ 0, & \text{otherwise}, \end{cases} \tag{2.3}$$

   s_i는 $\{m_j : j \in J_i^k\}$의 중위수로 정의된다.

로버스트성 가중값 r_i를 이용하여 적합된 국소선형회귀모형은 이상값의 영향을 덜 받는 로버스트 추정결과를 산출하게 되는데, Cleveland (1979)가 제안한 Lowess (Locally weighted regression)와 비교할 수 있는 매우 경쟁력 있는 로버스트 추정방법임이 Park (2004)의 모의실험을 통하여 밝혀졌다.

로버스트 가중값 r_i 값이 상대적으로 적은 관찰값은 이상값일 가능성이 높다고 할 수 있기 때문에 선형보간법에 참여를 시키지 않는 것이 정상적인 유사자료만의 생성을 보장하는 길이 될 것이다. 따라서 다음과 같이 수정된 선형보간법에 의한 유사자료의 생성방법을 제안할 수 있다.

1. 자료 $\{(X_i, Y_i)\}_{i=1}^n$가 주어지면 식 2.2에 정의된 로버스트성 가중값 r_i를 계산한다.
2. $r_i \geq \alpha_1$에 해당하는 자료 (X_i, Y_i)만을 대상으로 Hall과 Turlach의 선형보간법을 실시하여 유사자료를 생성한다.
3. $r_i \leq \alpha_2 (< \alpha_1)$에 해당하는 자료는 이상값일 가능성이 대단히 높다고 할 수 있기 때문에 원자료에서 제외시킨다.

상수 α_1과 α_2의 값을 적절하게 정하기 위해서는 가중값 r_i의 정확한 분포를 알아야 할 것이다. 그러나 r_i의 정확한 분포를 유도하는 작업이 아직 이루어지지 않았을뿐더러, 정확한 분포의 유도가 과연 가능한지 여부도 아직 답을 할 수 없는 상태이다. 그러나 여러 가지 경우를 고려하여 이루어진 수 많은 실험자료에서 얻어지는 결론으로는 $\alpha_1 = 0.1$, $\alpha_2 = 0.01$의 값이 가장 무난한 것으로 나왔다. 물론 좀더 세밀한 추가 확인이 필요하겠으나 모의실험의 결과를 볼 때 큰 문제없이 사용할 수 있는 상수값이라고 하겠다.

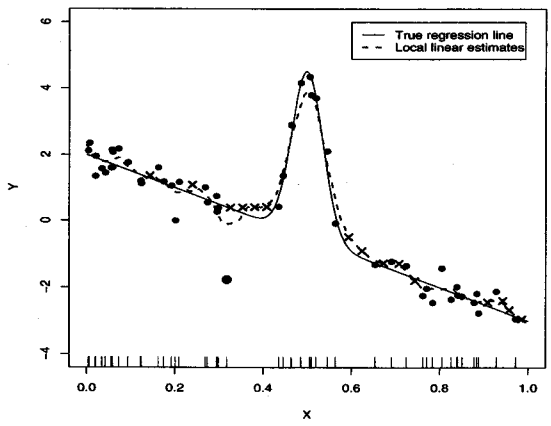

그림 2.1: 이상값을 포함시킨 그림 1.3의 자료에 대하여 수정된 선형보간법을 적용시켜 얻은 유사자료와 국소선형 추정량의 결과

따라서 이 논문에서 제안되는 수정된 선형보간법에 의한 유사자료 생성방법은 우선 로버스트성 가중값 r_i의 값이 0.01이하인 자료는 원자료에서 완전히 삭제시키며, 0.01이상 0.1이하인 자료의 경우에는 원자료에 남겨두기는 하지만 선형보간법에 포함을 시키지는 않고, 가중값 $r_i \geq 0.1$인 자료의 경우에만 유사자료의 생성과정에 포함시킴으로 해서 이상값에 의해서 왜곡되는 현상을 미연에 방지하는 안전판을 제공하는 것이다.

그림 2.1는 이상값을 포함시킨 그림 1.3의 자료를 수정된 선형보간법에 적용시켜 얻은 유사자료와 국소선형 추정량에 의하여 추정된 회귀곡선을 표시하고 있다. $x = 0.4$ 부근에서 생성된 4개의 유사자료들은 이상값의 영향을 전혀 받지 않은 매우 이상적인 형태임을 알 수 있는데, 이것은 이상값으로 포함된 $(X_{25}, Y_{25}) = (0.317, -1.766)$의 로버스트성 가중값이 $r_{25} = 0.09$으로 계산되어 유사자료의 생성과정에 포함되지 않았기 때문이다.

3. 모의실험

국소선형회귀모형의 추정량이 겪게 되는 자료 희소성 문제를 해결하기 위하여 Hall과 Turlach가 제안한 선형보간법이 이상값이 존재하는 경우에는 매우 왜곡된 결과를 낳을 수 있다는 사실은 그림 1.3을 통하여 확인할 수 있었다. 또한 이상값의 영향력을 감소시킬 수 있다고 제안된 수정된 선형보간법의 효과는 그림 2.1를 통하여 확인하였으나, 단지 한 예제의 결과에 불과한 것이므로 좀더 확실한 결론을 내리기 위하여 모의실험을 실시하였다.

모의실험에서 사용된 회귀함수는 식 (1.2)에 정의되어 있는 함수 $g(x)$이다. 이 회귀함수는 Seifert와 Gasser (1996), 그리고 Hall과 Turlach (1997)에서 사용된 회귀함수인데, 직선과 뾰족한 봉우리 모양의 곡선이 결합된 형태를 취하고 있어서 정확하게 추정하기가 쉽지 않은 함수임을 알 수 있다.

표본크기는 $n = 50$만이 사용되었고 독립변수의 설계점 X_i들은 Uniform$(0, 1)$의 분포에서 난수를 발생하여 얻은 후 정렬을 하여 $X_1 \leq X_2 \leq \cdots \leq X_n$의 관계를 만족하도록 하였다. 일단 X_i들이 주어지면 $S_i = X_{i+1} - X_i$, $i = 1, \ldots, n-1$가 계산되는데, 이것들 중에 가장 큰 값을 $S_l = \max_i S_i$이라 하자. 즉, l번째 구간의 길이가 가장 길다고 하자.

종속변수 Y_i는 $g(X_i) + \epsilon_i$로 정의되어지므로 오차항 ϵ_i의 분포에 따라 이상값이 포함될 가능성 및 그 정도가 변하게 되며, 다음의 5가지 상황이 고려되었다.

S1. $\epsilon_i \sim N(0, .5), i \neq l$. 그리고 $\epsilon_l \sim N(0, \sigma^2)$. 단, $\sigma = 2, 3$.

S2. $\epsilon_i \sim N(0, .5), i \neq l, l+1$. 그리고 $\epsilon_j \sim N(0, \sigma^2), j = l, l+1$. 단, $\sigma = 2, 3$.

S3. $\epsilon_i \sim .8N(0, .5) + .2N(0, 4)$, $i = 1, \ldots, n$.

S4. $\epsilon_i \sim .9N(0, .5) + .1N(0, 16)$, $i = 1, \ldots, n$.

S5. $\epsilon_i \sim$ Cauchy$(0, s)$, $i = 1, \ldots, n$. 단, $s = .1, .5$.

각 상황마다 자료가 주어지면 다음의 3가지 방법으로 자료를 보정한 후에 국소선형회귀모형의 추정량으로 회귀함수를 각각 추정하였다.

M1. 자료보정없이 원자료만을 이용한다.

M2. Hall과 Turlach의 선형보간법으로 생성된 유사자료를 원자료에 포함시킨다.

M3. 수정된 선형보간법으로 보정된 자료를 이용한다.

국소선형회귀모형의 추정량을 위해서 커널함수는 Epanechnikov 함수를 사용하였고, 띠폭은 점근적 MISE를 최소화시키도록 유도된 값인 $h = 0.0457$을 이용하였다.

수정된 선형보간법을 사용하기 위해서는 식 2.2에 정의된 로버스트성 가중값 r_i를 계산해야 하는데, 이를 위해 첨자집합 J_i^k를 정의하기 위한 상수 k의 값을 정해야 한다. 명목상 k의 역할은 매우 중요하지만 $k = 5$부터 $k = 11$까지의 모의실험 결과에 큰 차이가 나지 않는 것으로 봐서 실질적으로는 결과에 미치는 영향이 미미한 것으로 볼 수 있겠다. 이 논문에는 $k = 9$의 결과를 인용하였다.

주어진 자료에 대하여 각 방법의 ISE (Integrated squared error)는 $m = 400$개의 격자점 $t_1, \ldots, t_m$에서 다음과 같이 구해졌다.

$$\sum_{i=1}^{m} (\hat{g}_j(t_i) - g(t_i))^2 .$$

단, $\hat{g}_j(x)$, $j = 1, 2, 3$은 각각 방법 M1, M2, M3에 의하여 보정된 자료를 이용한 국소선형추정량을 의미한다.

이어서 방법 M1, M2, 그리고 M3의 효율성 비교는 $N = 1000$번의 모의실험 반복을 통하여 계산된 ISE의 평균으로 정의되는 MISE (Mean integrated squared error)의 값으로 이루어졌다. 모의실험 결과는 표 3.1에 보고되어 있다.

모든 경우에서 M1의 MISE는 비정상적으로 크게 계산되었음을 알 수 있는데, 이것은 국소선형회귀모형의 추정량이 가지고 있는 자료 희소성 문제의 심각성을 드러내는 것이라 하겠다. 따라서 독립변수의 설계점이 임의설계법 (random design)에 의하여 선택되며, 비교적 작은 크기의 띠폭을 사용하는 경우라면 반드시 선형보간법으로 생성된 유사자료를 원자료에 포함시켜야 한다는 것을 알 수 있다.

표 3.1: Hall과 Turlach의 선형보간법과 수정된 선형보간법의 효율성 비교

	M1	M2	M3
S1, $\sigma = 2$	104.91	0.3471	0.3231
S1, $\sigma = 3$	132.08	0.4800	0.3619
S2, $\sigma = 2$	1180.3	0.4627	0.4139
S2, $\sigma = 3$	1892.1	0.7391	0.5091
S3	322.17	0.4726	0.4388
S4	789.03	0.7565	0.4665
S5, $s = 0.1$	194.97	25.784	0.2096
S5, $s = 0.5$	4871.3	642.54	0.9578

또한 모든 경우에 있어서 수정된 선형보간법이 Hall과 Turlach의 선형보간법보다 더 효율적인 결과를 낳는다는 것을 알 수 있는데, 특히 이상값의 포함정도가 심한 Cauchy 분포의 경우에는 Hall과 Turlach의 방법이 국소선형 추정량의 안전판 역할을 거의 수행할 수 없음을 알 수 있다. 이상값의 포함정도가 심하지 않은 경우에도 거의 동일하거나 약간 더 좋은 추정결과를 보이고 있는 것을 볼 때, 수정된 선형보간법의 우수성이 잘 드러난 모의실험 결과라고 하겠다.

4. 결론

좋은 특성을 많이 가지고 있는 국소선형회귀모형 추정량이 자료가 희박한 구간에서는 매우 왜곡된 추정결과를 보이는, 이른바 자료 희소성 문제를 갖고 있다는 것은 잘 알려진 사실이다. 또한 이 문제에 대한 가장 일반적인 해법은 Hall과 Turlach의 선형보간법이며, 이 방법이 상당히 안정된 결과를 보이고 있다는 것도 잘 알려진 사실이다. 그러나 선형보간법의 이러한 특성은 자료에 이상값이 없는 이상적인 경우에만 적용되는 것이며, 만일 이상값이 자료에 포함되고 선형보간법의 기준값으로 사용된다면 크게 잘못된 추정결과를 얻을 수 있다는 사실은 아직 지적된 바가 없었다.

이 논문에서는 Hall과 Turlach가 제안한 선형보간법이 이상값이 존재하는 경우에는 왜곡된 결과를 낳을 수 있음을 인식하고 그 대안으로서 이상값의 영향을 감소시킬 수 있는 수정된 선형보간법을 제안하였다. 이상값의 영향을 감소시키는 방법으로는 Park (2004)가 제안한 로버스트성 가중값을 이용하여, 그 가중값이 극히 적은 값이 되는 자료를 선형보간법에 포함시키지 않는 방법을 사용하였다. 제안된 방법은 사용하기 무척 간단하며, 예제와 모의실험을 통하여 입증된 우수성을 비추어 볼 때 매우 실질적인 방법이라고 생각된다.

참고문헌

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association*, **74**, 829–836.

Fan, J. (1992). Design-adaptive nonparametric regression, *Journal of the American Statistical Association*, **87**, 998–1004.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies, *The Annals of Statistics*, **21**, 196–216.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Application*, Chapman & Hall/CRC, New York.

Hall, P. and Turlach, B. A. (1997). Interpolation methods for adapting to sparse design in nonparametric regression, *Journal of the American Statistical Association*, **92**, 466–476.

Park, D. (2004). Robustness weight by weighted median distance, *Computational Statistics*, **19**, 367–383

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

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*, Chapman & Hall/CRC, London.

[2007년 4월 접수, 2007년 7월 채택]

Robust Interpolation Method for Adapting to Sparse Design in Nonparametric Regression*

Dongryeon Park[1)]

ABSTRACT

Local linear regression estimator is the most widely used nonparametric regression estimator which has a number of advantages over the traditional kernel estimators. It is well known that local linear estimator can produce erratic result in sparse regions in the realization of the design and the interpolation method of Hall and Turlach (1997) is the very efficient way to resolve this problem. However, it has been never pointed out that Hall and Turlach's interpolation method is very sensitive to outliers. In this paper, we propose the robust version of the interpolation method for adapting to sparse design. The finite sample properties of the method is compared with Hall and Turlach's method by the simulation study.

Keywords: Linear interpolation, local linear regression model, outliers, pseudo data, robustness weights.

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

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