응용통계연구 제 6권 2호
1993년 9월 pp.227-235.

비모수적 커널교정과 구간추정[1)]

이 재창[2)], 전 명식[2)], 김 대학[3)]

요 약

순서쌍으로 주어진 자료 (x_i, y_i), $i=1,2,\cdots,n$ 들에 대한 독립변수와 관련된 추정은 회귀분석과는 달리 교정(calibration)이라고 불리워진다. 본 논문에서는 정규성 등과 같은 가정을 하지않고 비모수적인 커널방법을 이용하여 교정함수를 추정하고 추정된 교정함수의 붓스트랩 신뢰대를 이용한 독립변수의 구간추정을 제안하고자 한다. 교정과 커널방법에 대해 설명하였으며 독립변수의 추정에 대한 문헌적 고찰과 함께 붓스트랩 신뢰대에 대하여 첨언하였고 실제 자료를 통하여 다른방법과 비교,분석하였다.

1. 교정과 커널 회귀함수

1.1. 교정

교정(calibration)이란 서로 관련된 순서쌍으로 주어진 자료들 (x_i, y_i), $i=1,2,\cdots,n$ 에대해 쉽게 얻어진 종속변수 y값에 대한 독립변수 x값의 추정에 관련된 제반절차를 통칭한다. 이 때 y는 x에 비해 비교적 쉽게 얻을 수 있고 x의 관찰에는 y의 관찰보다 노력이나 경제적인 비용이 많이 들거나 x를 직접 관찰할 수 없는 경우이다. 이러한 독립변수 x의 추정문제는 측정하기 힘들거나 측정될 수 없는 양에 대한 정보의 제공측면과 다방면에로의 응용측면에서 최근 활발히 연구,사용되고 있다.

교정은 크게 두 단계로 나누어질 수 있다. 첫째단계는 회귀분석과 마찬가지로 교정함수(calibration function) $g(x)$를 추정하는 것이다. 둘째단계는 주된 단계로서 추정된 교정함수의 역함수를 찾는 일이다. 이 단계에서는 추정된 회귀함수가 종속변수의 구간추정이나 점추정에 이용되는 회귀분석과는 달리 추정된 교정함수의 역함수가 독립변수 x의 점추정과 구간추정에 이용된다. 추정된 교정함수가 x의 구간추정에 이용되기 위해서는 단순 역함수를 가져야 한다. 그러나 실제로 역함수의 정확한 함수형태는 거의의 이용되지 않고 있다.

1.2 커널 회귀함수

교정은 회귀분석과 밀접한 관계를 유지하고 있기 때문에 회귀함수의 추정방법은 교정에 직접적인 영향을 초래한다. 그러나 오차의 정규성과 회귀함수의 선형성을 전제로 추정된 회귀함수는 실제 자료의 경우 실용성이 적을수 있으며 특히 역함수가 고려되는 교정에서는 더 큰 오류

1) 본 연구는 한국과학재단 연구과제 KOSEF 911-0105-016-1의 연구 지원에 의한 결과임.
2) (136-701) 서울시 성북구 안암동 고려대학교 통계학과 교수
3) (608-738) 부산시 남구 우암동 부산외국어대학교 통계학과 조교수

를 초래할 수 있다.

최근 급속한 컴퓨터의 발달과 함께 고전적 추정량들의 단점을 보완한 커널 추정방법(kernel estimation)이 회귀함수의 추정에도 이용되고 있다. 회귀함수의 커널추정량 $\hat{g}_n(x : h)$는 Watson(1964)과 Nadayara(1964)에 의해 독립적으로 제안되었으며 그 형태는

$$\hat{g}_n(x : h) = \frac{\sum_i K(\frac{x-x_i}{h})y_i}{\sum_i K(\frac{x-x_i}{h})} \tag{1.1}$$

이고 여기서 $K(\cdot)$는 대칭인 임의의 확률밀도함수로서 커널이라고 불리워지고 h는 평활의 양을 좌우하는 평활계수(smoothing parameter)로서 표본의 수 n 이 커질수록 0 으로 가까이 가는 한편 nh 는 무한대로 수렴하는 성질을 만족한다. 평활계수의 선택문제는 많은 사람에 의해 연구되어 왔으며 좋은 추정량으로서의 다양한 성질들이 규명되었다. 또 이론적인 기준하에서의 평활계수의 선택에 관한 연구도 활발하게 진행되어 왔다. 최근, 자료자체에 근거한 평활계수의 선택방법이 계속 연구되고 있고 대표적인 방법으로서 Rudemo(1982)와 Bowman(1984)에 의해 시작된 교차타당성(cross-validation)방법과 Jhun(1988)에 의해 시도된 붓스트랩(bootstrap)방법 등이 있다.

2. 독립변수의 추정

2.1. 모수적추정

전통적인 회귀함수의 추정이 그러하듯이 교정에서의 점추정도 미지의 교정함수의 선형성과 오차항의 정규성을 가정하고 이루어졌다. 예로서

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$y_i = \alpha + \beta x_i + \gamma x_i^2 + \varepsilon_i \quad i = 1, 2, \cdots, n$$

등과 같은 모형을 들 수 있다. 여기서 ε_i 는 평균이 0이고 분산이 σ^2인 정규분포를 따르고 α, β 그리고 γ는 상수이다. 이러한 가정하에서 교정함수는 독립변수에 대한 종속변수의 회귀분석(direct regression)과 종속변수에 대한 독립변수의 역회귀분석(inverse regression)에 의해 점추정될수 있다. Krutchkoff(1967)는 역회귀분석에 의한 점추정을 제안하고 회귀분석에 비해 평균제곱오차의 관점에서 우위임을 보였고 Ali와 Singh(1981)은 회귀분석과 역회귀분석방법의 가중평균 추정치를 제안한 바 있다. 커널함수를 이용한 점추정은 추정된 커널회귀함수의 역함수를 이용함으로서 구해진다.

독립변수의 구간추정은 Scheffe(1973)에 의하여 시도되었다. Scheffe는 알려진 함수 $f_0, f_1, \cdots, f_K$ 와 미지의 상수 $\beta_0, \beta_1, \cdots, \beta_K$에 대해 다음과 같은 선형모형을 고려하였다.

$$g(x) = \sum_{k=0}^{K} \beta_k f_k(x) \tag{2.1}$$

이때 오차항의 분산 σ^2이 알려져 있고 회귀함수가 단조증가함수일 경우 Scheffe는 $g(x)$의 최소제곱추정량 $\hat{g}(x)$를 이용하여 $g(x)$에 대한 $100(1-\delta)\%$ 신뢰대 $\hat{g}(x) \pm W(x)$ 를 구하고 이 신뢰대를 $100(1-\alpha)\%$를 만족하도록 알려진 σ의 $z_{\alpha/2}$ 배를 더하여 새로운 관찰치 y^*에 대하여

독립변수의 신뢰구간 추정치 $[U_0^{-1}(y^*), L_0^{-1}(y^*)]$를 구하였다. 여기서 L_0와 U_0는

$$U_0(x) = \hat{g}(x) + W(x) + \sigma \cdot z_{\alpha/2}$$
$$L_0(x) = \hat{g}(x) - W(x) - \sigma \cdot z_{\alpha/2}$$

를 만족하는 단조증가함수이다. σ가 알려져있지 않은 경우는 추정되어야 하고 $W(x)$도 수정되어야 한다.

그러나 교정함수 $g(x)$는 자료와 가정된 모형으로부터 추정되기 때문에 X의 구간추정은 확률적 서술(Uncertainty Statement)과 함께 주어져야 한다(Scheffe, 1973). 이러한 확률적 서술과 관련된 확률에는 α와 δ가 수반된다. Scheffe는 이러한 확률해석을 "in the long run greater than or equal to"란 문장으로 표현하였다. 그의 관점은 허용한계(tolerance limit)와 그 개념을 공유하고 있다. 이 두 확률에 대해 간단히 설명하면, δ는 교정함수의 신뢰대와 관련된 확률이고 α는 미래의 각 관찰치의 오차에 관련된 불확실성의 정도로 해석될 수 있다.(관찰치는 오차와 함께 측정됨). 그래서 미지의 독립변수가 신뢰구간 추정치에 포함된다는 서술이 올바른 비율이 결국은 $1-\alpha$보다 크거나 같다는 의미를 가진다.

2.2 비모수적 구간추정

독립변수의 점추정이나 구간추정은 높은 정도와 함께 제공되어야 함은 당연한 일이다. 그러나 가정된 선형모형과 미지의 교정함수와의 차이는 상당한 영향을 미칠 수 있다. 많은 경우 교정문제는 보다 더 일반적인 상황에서 해결되어져야 한다. 구체적으로 관찰된 자료가

$$y_i = g(x_i) + \varepsilon_i, \quad i = 1, 2, \cdots, n \tag{2.2}$$

의 모형을 만족한다고 가정하자. 여기서 ε_i 는 평균이 0이고 분산이 σ^2인 독립인 오차확률변수이고 $g(\cdot)$의 함수형태는 가정되지 않았다.

독립변수의 비모수적 구간추정은 비모수적 회귀분석(nonparametric regression)방법에 의해 해결될 수 있다. Lechner외 2인(1982)은 핵탱크의 부피와 압력의 자료에 대하여 스플라인(spline)이라고 불리워지는, 구간에서 연속인 다항함수를 이용하여 통계적인 신뢰극한을 제공하는 비모수적 방법을 제안하였다. 또 Knafl(1984)은 새로운 비모수적 방법을 제안하였다. 본 연구에서는 커널추정방법을 이용하여 교정함수를 추정하고 또 이의 붓스트랩 신뢰대를 이용하여 독립변수의 구간추정을 하고자 한다.

2.3 붓스트랩 신뢰구간추정

2.3.1. 붓스트랩 신뢰대

통계적으로 관심있는 함수의 표본분포를 추정하는 붓스트랩 방법은 여러가지 장점을 가지고 있다. 그 중에서도 모수나 모수의 함수에 대한 신뢰영역의 제공은 여러 분야에 광범위한 응용성을 가지고 있다. 많은 경우에 있어서 신뢰영역은 표본분포의 대표본근사방법(large sample approximation)에 의해 진행되어 왔다. 그러나 이 방법은 하나 혹은 그 이상의 장애모수가 존재할 경우 장애모수의 추정과 관련되어 쉽게 해결되지 못하지만 붓스트랩 방법의 장점은 장애모수의 추정없이도 추축통계량(pivotal statistic)의 붓스트랩분포를 이용하여 붓스트랩신뢰영역을 제공해 주는 점이다. 커널회귀함수의 붓스트랩 신뢰영역은 교정에서의 독립변수의 구간추정에 바로 이용될 수 있다.

교정함수의 붓스트랩 신뢰대를 구하기 위하여 먼저 평활계수 h를 자료에 근거한 방법을 이용하여 적절히 선택한 다음 추축통계량

$$\hat{g}_n(x : h) - g(x) \tag{2.3}$$

의 붓스트랩 분포를 다음과 같은 순서로 구하고 $100 \cdot (1-\delta)\%$ 붓스트랩 신뢰대를 구한다.

고정된 h와 독립변수의 교정구간내에 있는 특정 x에 대해

i) 교정함수의 커널추정량 $\hat{g}_n(x : h)$ 를 구한다.

ii) $\tilde{\varepsilon}_i = y_i - \hat{g}_n(x_i : h)$를 구하고 중심화된 잔차 $\hat{\varepsilon}_i = \tilde{\varepsilon}_i - \sum_{i=1}^{n} \tilde{\varepsilon}_i / n, i = 1,2,\cdots,n$ 를 얻는다.

iii) 중심화된 잔차의 붓스트랩 표본 $\hat{\varepsilon}_i^*$, $i = 1,2,\cdots,n$ 을 얻는다.

iv) 자료 y_i의 붓스트랩 표본 $y_i^* = y_i + \hat{\varepsilon}_i^*$, $i = 1,2,\cdots,n$ 을 구한다.

v) y_i^*들에 의한 $\hat{g}_n^*(x : h) - \hat{g}_n(x : h)$ 를 계산한다. 여기서

$$\hat{g}_n^*(x : h) = \sum_i K(\frac{x-x_i}{h}) y_i^* \ / \sum_i K(\frac{x-x_i}{h}).$$

vi) iii)에서 v)까지를 B 번 반복하여 추축통계량의 붓스트랩분포를 얻는다.vii) 붓스트랩 분포의 $100(1-\delta)\%$에 해당되는 상,하위 백분위수 $z_{\delta_x''}$, $z_{\delta_x'}$ 를 얻는다.

교정구간내의 다른 x값에 대하여 iii)에서 vii)까지의 과정을 반복한다. 그리하여 교정함수 $g(x)$의 붓스트랩 신뢰대

$$\hat{g}_n(x : h) - z_{\delta_x''} \le g(x) \le \hat{g}_n(x : h) + z_{\delta_x'} \tag{2.4}$$

를 얻는다.

2.3.2. 붓스트랩 구간추정

독립변수에 대한 구간추정은, 이제 붓스트랩 신뢰대의 역함수 변환에 의해 추정할 수 있다. 그러나 2.1절에서 설명한 두 확률 α, δ를 고려하여야 한다. 먼저 주어진 δ의 확률을 만족하는 붓스트랩 신뢰대로부터 새로운 관찰치 y^*의 구간추정치 $[B_l, B_r]$를 구한다. 두번째 확률 α를 만족시키기 위하여 추축통계량

$$\hat{g}_n^{-1}(y^* : h) - g^{-1}(y^*) \tag{2.5}$$

을 고려하여 B_l과 B_r각각에 대한 추축통계량 (2.5)의 붓스트랩분포를 유도하고 $100(1-\alpha)\%$ 백분위수 $z_{\alpha'}, z_{\alpha''}$를 구한다. 이제 $100(1-\delta)\%$ 붓스트랩신뢰대로부터 구한 구간추정치 $[B_l, B_r]$에 $100(1-\alpha)\%$ 백분위수 $z_{\alpha'}, z_{\alpha''}$를 활용하여 독립변수의 붓스트랩구간추정치 $[B_l + z_{\alpha'}, B_r + z_{\alpha''}]$를 얻는다.

3. 자료분석의 예

제안된 커널방법을 이용한 교정의 독립변수에 대한 구간추정의 타당성을 알아보기 위하여 교정분야에서 사용되었던 압력과 부피의 자료를 고려하였다.

<표 1> 압력과 부피의 자료

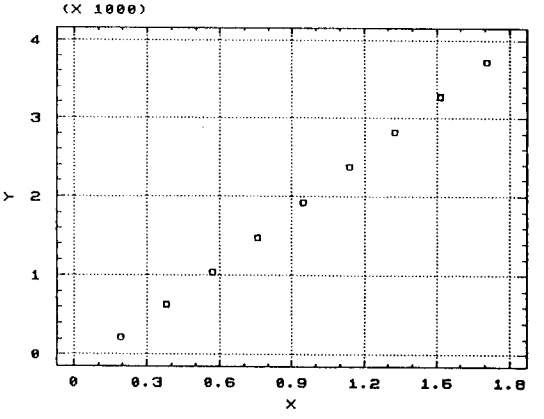
x	y	x	y
.18941	215.250	.94746	1921.85
.18949	218.281	1.13642	2372.04
.37880	632.621	1.13646	2374.43
.37884	627.141	1.13665	2377.76
.37884	628.711	1.32640	2819.47
.56840	1034.051	1.32640	2815.70
.56843	1033.341	1.51523	3263.50
.75755	1469.111	1.51528	3261.94
.75767	1474.111	1.51561	3268.54
.75769	1475.241	1.70525	3711.64
.94740	1924.520		

(x의 단위 : 1000 리터, y의 단위 : 파스칼)

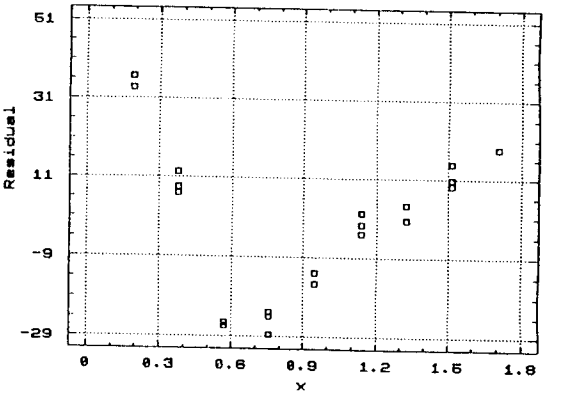
이 자료는 핵의 보안문제에 관련된 것으로 핵 저장탱크의 부피와 압력이 측정되어 <표 1>에 나타나 있다. 이들은 일정한 간격으로 측정되었으나 부피는 핵 저장탱크의 꼭대기 부분에서 측정된 미분압력(differential pressure)에 의하여 간접적으로 측정된 것이다. 이 자료의 경우 부피 x 에서의 압력을 $g(x)$라 하면 관측된 압력 y는 식 (2.2)에 의해 모형화될 수 있다. 그러나 탱크하부구조의 복잡성은 $g(x)$에 대한 모형설정을 불가능하게 한다.

3.1 자료의 요약과 커널추정

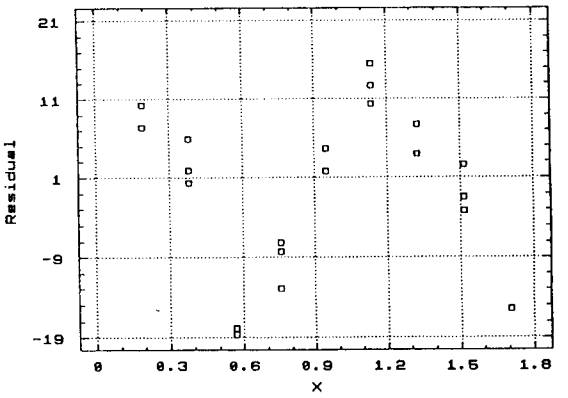
위의 자료를 <그림 1>에 나타내었다. <그림 1>을 보면 압력과 부피사이에는 선형성의 관계가 있는 것처럼 보이나 교정함수의 선형성이나 이차함수를 가정하여 구한 잔차들을 그린 <그림 2>와 <그림 3>을 보면 이들의 가정은 특성경향을 나타내어 만족스러운 결과를 제공하지 못함을 알 수 있다.



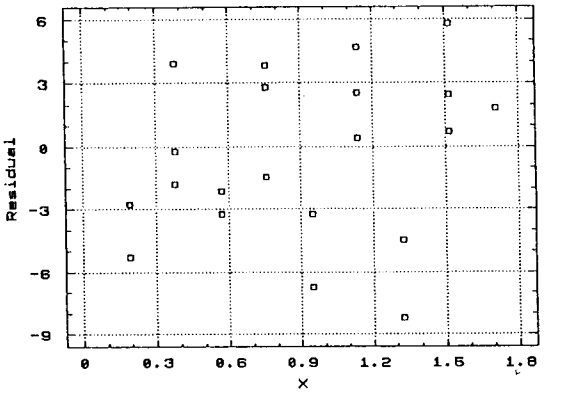
<그림 1> 자료의 산점도



<그림 2> 잔차그림(선형)



〈그림 3〉 잔차그림(이차함수)

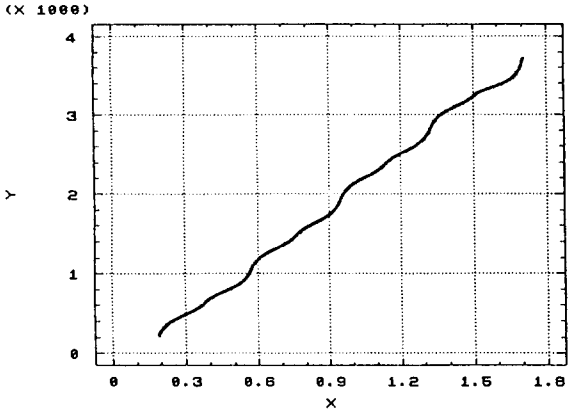


〈그림 4〉 잔차그림(커널)

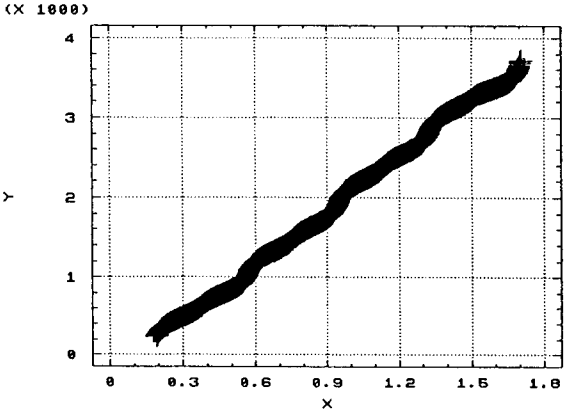
만족스럽지 못한 교정함수의 형태에 대한 가정을 배제하여 커널추정방법을 적용하여 보았다. 이때 평활계수의 선택방법으로 Kim(1990)에 의한 붓스트랩 방법을 적용하였고 Epanechnikov (1969)에 의해 제안된 커널

$$K(x)=\begin{cases} 0.75(1-x^2) & |x|\le 1 \\ 0 & |x|>1 \end{cases}$$

을 사용하였다. 또한 평균적분제곱오차를 선택기준으로 하여 100번의 붓스트랩 반복에 의하여 평활계수 h = 0.19 를 얻었다. 선택된 평활계수 0.19 에 의한 교정함수의 커널추정량은 <그림 5>에 나타내었고 <그림 4>에는 커널추정에 의한 잔차를 도시하였다. 이들 그림에서 알 수 있듯이 교정함수의 선형성이나 이차함수를 가정했을 때에 나타난 잔차들의 특성경향은 사라졌다.



〈그림 5〉 커널추정치



〈그림 6〉 붓스트랩신뢰대

3.2 독립변수의 붓스트랩구간추정

2장에서 설명한 붓스트랩 신뢰대는 교정함수의 동시신뢰대는 아니지만 교정분야에의 실용가능성을 보여주고 있으며 붓스트랩 동시신뢰대를 위해서는 수리적으로 조금은 복잡한 $|\hat{g}_n(x)-g(x)|$ 의 최대상계노움(supremum norm)으로 대신하면 된다.

붓스트랩 신뢰대를 얻기 위하여 교정구간을 1000등분하고 다른방법과 비교하기 위하여 δ를 0.1로 택하여 500번의 붓스트랩반복을 통하여 얻은 결과를 <그림 6>에 나타내었다.

종속변수의 새로운 26개의 자료들에 대하여 독립변수의 점추정량은 선형보간법(linear interpolation)에 의하여 얻을 수 있고 독립변수의 구간추정은 $100(1-\delta)\%$ 붓스트랩 신뢰대로부터 $100(1-\alpha)\%$의 확률을 만족하는 구간추정치를 얻을 수 있다. δ와 α가 각각 0.1일때의 결과가 <표 2>에 나타나 있다.

<표 2> 붓스트랩신뢰구간과 점추정

Y값	하한	점추정	상한	폭	비
229.1770	.1887	.1911	.1933	.0044	.3931
367.9880	.2258	.2285	.2311	.0053	.3367
506.8000	.3104	.3149	.3188	.0084	.6089
645.6110	.3824	.3847	.3872	.0048	.4508
784.4220	.4644	.4686	.4725	.0081	.5440
923.2330	.5410	.5431	.5452	.0042	.4349
1062.0400	.5716	.5731	.5748	.0031	.4419
1200.8600	.6074	.6097	.6122	.0048	.4247
1339.6700	.6882	.6922	.6953	.0071	.7638
1478.4800	.7572	.7597	.7619	.0047	.8734
1617.2900	.8252	.8287	.8329	.0078	.7990
1756.1000	.9049	.9072	.9096	.0046	.4048
1894.9100	.9403	.9419	.9434	.0031	.4411
2033.7200	.9665	.9687	.9706	.0042	.4341
2172.5300	1.0239	1.0282	1.0314	.0075	.6725
2311.3500	1.1098	1.1128	1.1159	.0060	.8482
2450.1600	1.1645	1.1677	1.1706	.0061	.8119
2588.9700	1.2513	1.2548	1.2576	.0063	.5704
2727.7800	1.3074	1.3095	1.3114	.0040	.4419
2866.5900	1.3329	1.3344	1.3357	.0028	.3625
3005.4000	1.3718	1.3749	1.3774	.0057	.5091
3144.2100	1.4557	1.4598	1.4634	.0077	.8632
3283.0200	1.5225	1.5280	1.5334	.0109	1.8779
3421.8400	1.6384	1.6429	1.6461	.0077	.7647
3560.6500	1.6863	1.6882	1.6899	.0036	.2746
3699.4600	1.7027	1.7041	1.7053	.0027	.2955

위의 표에는 Knafl(1984)등이 사용한 26개의 새로운 y값과 커널방법에 의한 점추정값, 붓스트랩신뢰구간의 상한,하한값 그리고 붓스트랩신뢰구간의 폭도 함께 나타나 있다. 또한 비교를 위하여 Knafl에 의해 추정된 신뢰구간의 폭을 분모로, 커널과 붓스트랩방법에 의한 신뢰구간의 폭을 분자로 한 비를 마지막 열에 나타내었다. 비의 결과를 보면 한 관찰치를 제외하고는 커널방법에 의한 붓스트랩신뢰구간의 폭이 상당히 짧아졌음을 알 수 있다. 신뢰구간의 폭이 줄어든 것은 교정함수의 형태를 가정하지 않은 커널방법의 우월성을 보여주는 것이고 미지의 교정함수의 독립변수에 대한 신뢰구간을 붓스트랩신뢰영역으로 잘 추정했음을 의미한다고 볼 수 있겠다.

4. 결 론

이상에서 살펴본 바와 같이 회귀분석이나 교정등에서 전통적으로 사용되어온 미지의 함수에 대한 선형성이나 오차확률변수의 정규성등의 가정은 실제자료의 경우에 잘 맞지않으며 이런 강한 가정들을 배제한 커널방법에 의한 비모수적 교정을 수행함이 바람직한 방법으로 생각된다. 또 붓스트랩방법에 의한 교정함수의 신뢰영역의 제공은 흥미로운 결과를 가져다 주었다.

참 고 문 헌

[1] Ali, M. A. and Singh, N.(1981), "An alternative estimator in inverse linear regression," *Journal of Statistical and Computational Simulation*, **14**, 1-15.

[2] Bowman, A. (1984), "An alternative method of cross-validation for the smoothing of density estimates," *Biometrika*, **71**, 353-360.

[3] Epanechnikov, V. A. (1969), "Nonparametric estimation of multidimensional probability density," *Theor. Probab. appl.*, **14**, 153-158

[4] Jhun, M. (1988), "Bootstrapping density estimates," *Communications in Statistics, Theory Methods*, **17**, 61-78.

[5] Kim, D. (1990), *Bootstrap choice of bandwidth for nonparametric kernel regression*., Ph.D dissertation, Dept. of Statistics, Korea University.

[6] Knafl, G., Spiegelman, C., Sacks, J. and Yilvisaker, D. (1984), "Nonparametric calibration," *Technometrics*, **26**, 233-241.

[7] Krutchkoff, R. G. (1967), "Classical and inverse methods of calibration," *Technometrics*, **9**, 525-539.

[8] Lechner, J. A., Reeve, C. P. and Spiegelman, C. H (1982), "An implementation of the Scheffe's approach to calibration using spline function illustrated by a pressure volume calibration," *Technometrics*, **24**, 229-234.

[9] Nadayara, E. A. (1964), "On estimating regression," *Theor. Probab. appl.*, **9**, 141-142.

[10] Rudemo, M. (1982), "Empirical choice of histograms and kernel density estimator," *Scandinavian Journal of Statistics*, **9**, 65-78.

[11] Scheffe, H. (1973), "A statistical theory of calibration," *Annals of Statistics*, **1**, 1-37.

[12] Watson, G. S. (1964), "Smooth regression analysis," *Shankya, Ser A.*, **26**, 359-372.

Nonparametric Kernel Calibration and Interval Estimation[1)]

Jae Chang Lee[2)], Myoungshic Jhun[2)], Daehak Kim[3)]

Abstract

Calibration relates the estimation of independent variable which requires more effort or expense than dependent variable does. It would be provided with high accuracy because a little change of the result of independent variable can cause a serious effect to the human being. Usual statistical analysis assumes the normality of error distribution or linearity of data. It is desirable to analyze the data without those assumptions for the accuracy of the calibration.

In this paper, we calibrated the data nonparametrically without those assumptions and derived confidence interval estimate for the independent variable. As a method, we used kernel method which is popular in modern statistical branch. We derived bootstrap confidence interval estimate from the bootstrap confidence band.

1) This research was supported by the Korean Science & Engineering Foundation Grant KOSEF 911-0105-016-1.

2) Department of Statistics, Korea University, Seoul, 136-701, Korea

3) Department of Statistics, Pusan University of Foreign Studies, Pusan, 608-738, Korea.