KJHM Vol. 16, No. 3, 2003, 95-100

통계학의 비모수 추정에 관한 역사적 고찰

서경대학교 소프트웨어학과 **이승우**

Abstract

The recent surge of interest in the more technical aspects of nonparametric density estimation and nonparametric regression estimation has brought the subject into public view. In this paper, we investigate the general concept of the nonparametric density estimation, the nonparametric regression estimation and its performance criteria.

0. 서론

통계학은 자료의 정리 및 요약 그리고 특성치 산출에 의한 자료의 분석을 할 수 있는 기술통계학과 확률이론을 기초로 한 모수 통계 및 비모수 통계로 구분할 수 있는 추측통계학으로 구분된다. 모수 통계는 모집단의 분포에 관한 특별한 가정 하에 통계적 추론을 시행하며 모집단의 특성을 대변하는 특성치인 모수에 관한 추론을 취급하고 있다. 그러나 현실적으로 모집단에 대한 가정이 충족된다고 볼 수 없는 상황이거나 모집단의 분포에 관한 특별한 가정을 필요로 하지 않은 경우, 즉 모수를 알 수 없거나 필요로 하지 않은 상황에서 통계적 추론을 시행할 수 있는 비모수 통계학이 있다.

비모수 통계학은 가정이 없이 통계적 추론을 하므로 상당히 일반적이고, 양적 관측값으로 표현할 수 없는 경우 순위를 사용함으로서 실제적이며, 특이점이 있는 경우 모수적 방법보다 검정력의 손실을 피할 수 있는 장점이 있다.

또한 모집단의 분포가 불투명하여 모수통계방법으로 적용한다면 큰 오차가 발생할 가능성이 예견될 경우, 즉 관측값을 측정하는 과정에서 基數的(cardinal) 변량 그리고 比率(ratio)과 같은 수량적 자료를 이용하는 모수적 통계방법보다는 수량적 자료뿐만 아니라 名目的 자료(nominal, categorical data) 그리고 순위를 나타내는 序數的(ordinal) 변량만을 가지고도 통계적 추론을 전개할 수 있는 비모수적 통계방법을 사용함으로서 강력한 추론방안을 제시해 준다. 한편, 비모수 함수 추정은 통계학의 한 부류로 급속하게 발전하고 있으며, 비모수 밀

도 추정은 1960년대의 통계학 연구의 주류였고 비모수 회귀 추정에 기초를 제공하고 있다.

1. 비모수 밀도 추정

확률변수 X의 행동을 묘사하기 위한 기본적인 특성으로서 확률밀도함수를 사용한다. 추세곡선과 곡면을 추정하는데 있어서 확률밀도함수는 다양한 의미로 많은 정보를 우리에게 부여해 줌으로서 자료분석에 있어서 도움을 주고 있다. 그러나 현실적으로 자료분석 및 실질적인 연구에서 확률변수 X에 대한 확률밀도함수를 정확히 알 수는 없다. 그러나 미지의 확률밀도함수 f는 서로 독립이고 동일한 분포로 이루어진 확률변수들로서 가정된 n개의 관측값 $\{X_i\}_{i=1}^{n}$의 집합에서 생성된다. 이 관측값들을 근거로 확률밀도함수를 추정하는 것이 우리의 목적이다.

확률밀도함수를 추정하는 방법으로서 모수적 접근방법과 비모수적 접근방법이 있다. 곡선을 추정하기 위한 모수적 접근방법은 모수만을 알 수 없기 때문에 확률밀도 추정은 모수 추정이라고 할 수 있다. 비모수적 접근방법은 근본적으로 확률밀도함수 곡선의 전체를 locally 또는 globally하게 추정하는 것이다. 이 비모수적 접근방법은 true density function의 형태나 class에 대한 정확한 정보를 전혀 가지고 있지 못하다. 그래서 비모수적 접근방법을 요약한다면 다음과 같이 언급할 수 있다. "Let the data speak for themselves."

가장 일반적으로 사용되어진 비모수 밀도 추정은 kernel을 이용한 kernel density estimation이다. Kernel estimator의 기본개념은 Rosenblatt에 의하여 최초로 소개되어졌고 그후 Pazen에 의하여 구체적으로 고안되어졌다.

$X_1, X_2, \cdots, X_n$은 확률변수 X로서 서로 독립이고 동일한 분포로 이루어진 확률변수들이라고 가정할 때, 확률분포함수 $F(x)=P[X\leq x]=\int_{-\infty}^{x} f(y)\,dy$은 연속이고 $f(x)$는 확률밀도함수이다. 임의의 주어진 점 x에서 분포함수 $F(x)$의 추정으로서, 다음과 같은 표본분포함수를 정의한다.

$$F_n(x)=\left(\frac{1}{n}\right)(X_1, X_2, \cdots, X_n \text{ 중에서 } x\text{보다 작은 관측값의 개수})$$

확률밀도함수의 여러 가지 다양한 추정치 들 중에서 평활폭 h를 사용하는 표본분포함수를 정의하면 다음과 같다.

$$f_n(x)=\frac{F_n(x+h)-F_n(x-h)}{2h}=\frac{1}{2hn}\{(x-h, x+h) \text{ 안에 존재하는 } X_i\text{의 개수}\}$$

여기서 평활폭 h는 양의 실수이며 평활폭 h의 선택방법이 중요한 문제이다. h는 n의 함

수이며 h는 0으로 n은 ∞로 접근하는 경향이 있다.

한편, 위에서 정의한 추정량을 연구하기 위해서 kernel을 이용한다. Kernel은 함수로서 Uniform, Triangle, Epanechnikov, Quartic, Triweight, Gaussian, Cosinus 등과 같이 다양하게 정의된다.

위에서 제시된 추정량은 표본분포함수 상에서 weighted average로서 사용된다.

$$f_n(x)=\int_{-\infty}^{\infty}\frac{1}{h}K\left(\frac{x-y}{h}\right)dF_n(y)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)$$

여기서 weight function $K(z)$는 kernel 함수이며 $K(z)$는 연속이고 대칭이며 유계(bounded)이다. 그리고 kernel 함수는 $\int K(z)\,dz=1$을 만족한다.

그러므로 커널밀도추정(kernel density estimate)은 다음과 같이 정의한다.

$$\hat{f}_h(x)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)$$

2. 비모수 회귀 추정

자연, 사회, 의학 등의 여러 학문 연구 분야에서 또는 일상 생활에 적용된 둘 또는 두개이상의 변수들의 상호 관련성을 분석해야 할 경우가 많이 있다. 그러므로 통계분석에서 큰 비중을 차지하는 변수들간의 관계에 대한 연구와 더불어 변수들간의 상호 관련성을 수학적인 함수의 형태로서 표현하여 변수들간의 관련성을 규명하고 한층 더 나아가 한 변수로부터 다른 변수들간의 변화를 예측하고자 한다.

우리의 생활 주변에는 독립 변수와 종속 변수 사이의 함수 관계를 알아내려는 수없이 많은 문제들이 있으며, 이런 관계를 통하여 실험 자료(data)를 분석하여 어느 정도 규명할 수 있다면 귀중한 정보를 우리에게 제공할 수 있을 것이다. 이처럼 변수들간의 함수관계를 추구하는 통계적 방법을 회귀분석(regression analysis)이라고 한다.

서로 독립이고 동일한 분포로 이루어진 확률변수들로 생성된 자료 $\{(X_i, Y_i)\}_{i=1}^{n}$를 고려해보자. 회귀분석은 설명변수 X와 반응변수 Y간에 일반적인 관계식인 함수로 표현된다. 즉 설명변수 $X=x$가 주어졌을 때 Y의 기대값인 확정적인 요소(deterministic elements), $m(x)=E(Y|X=x)$와 확률적인 요소(stochastic element)인 $Var(Y|X=x)=\sigma^2$로 구성되어있다. 그리고 $E(Y)<\infty$를 만족해야한다.

만약 결합밀도함수 $f(x,y)$가 존재하면 $m(x)$는 다음과 같다.

$$m(x)=\int yf(x,\ y)dy/f(x)$$

여기서 $f(x)=\int f(x,\ y)dy$는 X의 주변확률밀도함수이다.

회귀분석을 하는 목적은 확정적인 요소 $m(x)=E(Y|X=x)=\int yf(x,y)dy/f(x)$인 conditional expectation의 추정치를 찾는 것이 목적이다. 한편, multiplicative kernel을 이용한 결합밀도함수 $f(x,\ y)$는 다음과 같이 추정할 수 있다.

$$\hat{f}_{h_1,h_2}(x,\ y)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_1}K\left(\frac{x-X_i}{h_1}\right)\frac{1}{h_2}K\left(\frac{x-Y_i}{h_2}\right)$$

그리고 $m(x)$의 추정치인 $\widehat{m_h}(x)$는 다음과 같다.

$$\widehat{m_h}(x)=n^{-1}h^{-1}\sum_{i=1}^{n}\frac{1}{\hat{f}_h(x)}K\left(\frac{x-X_i}{h}\right)Y_i$$

여기서 함수 $\hat{f}_h(\cdot)$는 X의 주변확률밀도함수의 Rosenblatt-Parzen kernel density estimator이다.

Weight sequences, $\{W_{hi}(x)\}_{i=1}^{n}$로서, $W_{hi}(x)=\frac{1}{\hat{f}_h(x)}\frac{1}{h}K\left(\frac{x-X_i}{h}\right)$라 할 때, 비모수 회귀 추정의 일반적인 형태는 다음과 같다.

$$\widehat{m_h}(x)=n^{-1}\sum_{i=1}^{n}W_{hi}(x)\,Y_i$$

Weight sequences, $\{W_{hi}(x)\}_{i=1}^{n}$는 h에 의하여 조절되어지며 관측값 Y_i들을 평균한 값에 의하여 통제된다. 이때, 평활폭 또는 smoothing parameter인 h를 선택하는 방법이 가장 중요한 문제이다. 한편, $m(x)$의 추정치의 여러 가지 형태를 고려해보자. 위의 추정치의 형태는 다음과 같이 표현할 수 있으며 Nadaraya-Watson 추정량이라고 한다.

$$\widehat{m_h}(x)=\frac{n^{-1}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-X_i}{h}\right)Y_i}{n^{-1}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-X_i}{h}\right)}$$

관측값 X들은 거의 일정한 간격(거리)에서 취해지고 임의의 구간에 존재하는 점들의 equidistant grid를 형성한다. 일반적인 성질의 손실 없이 관측값 X들은 단위 구간 [0, 1]에서 취한다. 즉, $\max_i|X_i-X_{i-1}|=O(n^{-1})$일 때 $x\in(X_{i-1},\ X_i]$에서 $\hat{f}(x)=$

$[n(X_i - X_{i-1})]^{-1}$로서 주변확률밀도함수를 이용할 수 있다. 그러므로 $m(x)$의 추정치의 또 다른 형태로서, 다음은 Priestly와 Chao(1972) 및 Benedetti(1977)에 의하여 연구되었다.

$$\widehat{m_h}(x) = h^{-1}\sum_{i=1}^{n}(X_i - X_{i-1})K\left(\frac{x - X_i}{h}\right)Y_i$$

$m(x)$의 추정치의 또 다른 형태로서, $\{s_i\}_{i=0}^{n}$이 $s_0 = 0$, $s_{i-1} \le X_i \le s_i$ $(i = 1, \cdots, n)$, $s_n = 1$을 만족할 때 다음은 보다 작은 오차(error)들이 생성되며 Gasser와 Müller(1979)에 의하여 제시되어졌다.

$$\widehat{m_h}(x) = h^{-1}\sum_{i=1}^{n}\int_{s_{i-1}}^{s_i} K\left(\frac{x - u}{h}\right)du\, Y_i$$

많은 실험에서 설명변수 $\{X_i\}_{i=0}^{n}$는 구간 $[a, b]$ 위에서 equidistributed 되어진 상태로 취해졌으며, 일반적인 성질의 손실없이 $[a, b] = [0, 1]$을 만족한다. 그리고 $\hat{f}(x) = I_{[0,1]}$로서 주변확률밀도함수를 사용하며 $[0,1]$ 위에서 균일분포(uniform distribution)의 밀도함수이고 $X_i = i/n$을 가정한다.

또한 Priestley와 Chao 및 Gasser와 Müller에 의하여 위에서 제시된 두 개의 추정량을 다시 쓰면 다음과 같다. 그리고 이 두 추정식은 회귀곡선을 추정하는데 사용되며 이 두 추정식을 확장하면 회귀곡면을 추정하는데 사용될 수 있다.

$$\widehat{m_h}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)Y_i$$

$$\widehat{m_h}(x) = \frac{1}{h}\sum_{i=1}^{n}\int_{s_{i-1}}^{s_i} K\left(\frac{x - u}{h}\right)du\, Y_i \ , \ s_i = \frac{1}{2}(X_i + X_{i+1})$$

3. Performance criteria

통계학에서 추정량 $\hat{\theta}$와 target θ 간에 근접성을 측정하고자 평균제곱오차(mean squared error; MSE)의 크기를 사용한다. MSE는 분산과 편의(bias)의 자승으로 간단하게 표현된다.

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = Var(\hat{\theta}) + (E(\hat{\theta}) - \theta)^2$$

그러므로 MSE를 최소화하는 추정량 $\hat{\theta}$을 선택할 수 있다.

임의의 $x \in \Omega$에서 함수 $g(x)$의 추정량으로서 $\hat{g}(x; \cdot)$를 고려하자. $MSE\{\hat{g}(x_0; \cdot)\}$을

최소화함으로서 고정된 점 x_0에서 $g(x_0)$을 추정할 수 있다.

정의역 Ω 위에서 g를 추정하기 위한 바람직한 방법으로서 함수들 $\hat{g}(\cdot;\cdot)$와 g 간에 거리를 globally하게 측정하는 error criterion을 사용하여 함수 $\hat{g}(x;\cdot)$을 추정하자. 그러한 error criterion은 integrated squared error(ISE)이며 다음과 같이 정의한다.

$$ISE\{\hat{g}\} = \int \{\hat{g}(x;\cdot) - g(x)\}^2 dx$$

이것은 $\hat{g}(\cdot;\cdot)$와 g의 L_2거리의 제곱으로 표시할 수 있다. ISE는 주어진 모든 관측값들에 의하여 globally하게 측정하기 때문에 유일하게 적용됨으로서, ISE는 다른 조건에서 주어진 관측값들에는 적용할 수 없다. 그러므로 이러한 단점을 극복하기 위해서 Rosenblatt (1956)에 의하여 최초로 사용되어진 mean integrated squared error(MISE)는 다음과 같다.

$$MISE\{\hat{g}\} = E(ISE\{\hat{g}\}) = E\int \{\hat{g}(x;\cdot) - g(x)\}^2 dx = \int MSE\{\hat{g}(x;\cdot)\} dx$$

MISE criteria는 커널밀도함수와 회귀함수 m을 추정하는 데 사용될 수 있다.

참고 문헌

1. Eubank, R.L., *Spline Smoothing and Nonparametric Regression*, Marcel Dekker, New York, 1988.
2. Härdle, W., *Applied Nonparametric Regression*, Cambridge University Press, Cambridge, 1990.
3. Härdle, W., *Smoothing Techniques with Implementation in S*, Springer-Verlag, New York, 1991.
4. Silverman, B.W., *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York, 1986
5. Wand, M.P., Jones, M.C., *Kernel Smoothing*, Chapman and Hall, 1995.