General cross-validation과 Least squares method에 의한 비매개변수적 회귀모형의 특성

조성진[1], 문영일[2], 황성환[3], 박대형[4], 권현한[5]

1. 서론

어떤 현상이 변수들의 인과관계에 의하여 나타날 때 그 관계를 규명하는 것은 공학적으로 중요한 일이며, 이를 위하여 사용되는 방법들 중의 하나가 회귀분석이다. 수자원 분야에서 흔히 사용되는 방법들은 인자들간의 관계를 매개변수를 통하여 나타내게 되는 매개변수적 해석방법이다.

매개변수적 회귀모형과 구별되는 비매개변수적 회귀모형은 주어진 자료의 특성으로부터 잡음(noise)을 제거 또는 감소시킬 수 있으며, 자료가 지니는 특성에 보다 근접하는 회귀모형을 구할 수 있다는 장점을 지닌다. 비매개변수적 회귀분석은 매개변수적 회귀분석이 해석하기 어려운 자연계의 이질적이고, 다중변수, 시간과 공간적인 변수를 지니게 되는 자료들에 대한 유용한 해석방법이라고 할 수 있다(문영일 등, 2000).

Kernel regression은 대표적인 비매개변수적 회귀분석으로서 주어진 자료에 가중치를 부여함으로써 회귀모형을 산정하게 된다. Kernel regression의 중요한 인자 중 하나인 광역폭(bandwidth)을 결정하는 방법은 일반적으로 사용되는 Least squares method나 General cross-validation 등이 있으며, Bandwidth 선택방법에 따른 회귀모형의 특성을 비교하기 위하여 임의의 함수에 잡음(noise)을 첨가하고, Kernel regression을 실시, 비교하였다.

2. Kernel regression

매개변수적 방법과 같이 f에 대한 형태를 미리 설정하지 않고, 함수 f는 어떤 요건을 만족시키는 함수군에 속한다고 가정하여 자료로부터 원하는 지점의 회귀값을 추정하는 것이 비매매변수적 회귀모형의 기본원리이다.

Kernel regression estimator의 가장 기본적인 형태를 나타내면 다음과 같다.

$$f(x) = \frac{\sum_{j=1}^{n} K\left(\frac{(x-x_j)}{h}\right) y_j}{\sum_{j=1}^{n} K\left(\frac{(x-x_j)}{h}\right)} \tag{1}$$

여기서, h는 광역폭(bandwidth), n는 자료개수, (x_i, y_i)는 주어진 관측자료이며, x는 추정하고자하는 값, $K(\cdot)$는 Kernel 함수이다.

1), 3), 4), 5) 서울시립대학교 토목공학과 박사과정
2) 서울시립대학교 토목공학과 조교수

2.1 Kernel function

kernel regression에 있어서 kernel 함수는 관측값에 가중치로 작용하여 bandwidth내의 관측값으로부터 임의의 지점의 추정치를 찾아주는 역할을 한다. 관측값에 대한 가중치는 Kernel 함수의 모양에 의해 결정이 되며, 그 모양에 따라 여러 가지 종류가 있다.

kernel 함수의 일반적인 특징은 식 2~3과 같다.

$$\int_{-\infty}^{\infty} K(u)du = 1 \tag{2}$$

$$\int_{-\infty}^{\infty} uK(u)du = 0 \tag{3}$$

$$\int_{-\infty}^{\infty} u^2K(u)du = \alpha \text{ (} \alpha \text{는 0이 아닌 상수)} \tag{4}$$

2.2 Bandwidth

바람직한 회귀모형을 추정하기 위해서 자료가 나타내는 일련의 신호(signal)를 찾아내야 하는데 일반적으로 이 신호(signal)는 매끄럽다고 할 수 있다. 즉, 바람직한 회귀곡선을 "잔차의 제곱(오차)이 작으면서 매끄러운 곡선이다." 라고 할 수 있는 것이다(강근석, 김충락, 1999).

이 개념을 식으로 나타내면 다음과 같다.

i : $\Sigma\,[y_i - f(x_i)]^2$: 적합도 (goodness of fit)

ii : $\int_a^b f''(x)^2 dx$: 매끄러움 정도 (smoothness, roughness)

i 과 ii를 작게 하는 것이 올바른 회귀분석이라고 할 수 있다.

균형적인 조율을 위해서 , $0<q<1$ 을 가정하고, 식을 만들면 다음 식 5와 같다.

$$(1-q)\sum_{j=1}^{n}[y_i - f(x_i)]^2 + q\int_a^b [f''(x)]^2 dx \tag{5}$$

여기서, $\frac{q}{(1-q)} = \lambda$라고 하면 위 식은 다음 식 6과 같이 변형될 수 있다.

$$S(\lambda) = \sum_{j=1}^{n}[y_i - f(x_i)]^2 + \lambda\int_a^b [f''(x)]^2 dx \tag{6}$$

λ는 잔차제곱과 매끄러운 정도의 균형을 나타내는 값으로 bandwidth를 의미한다. λ가 아주 작으면, 들쭉날쭉해지고, λ가 아주 커지면 직선에 가까워진다. 즉, Kernel regression 에 있어서 bandwidth h로 사용되는 λ는 Kernel 함수의 광역폭을 의미하며, 이 크기에 따라서 관측값에 대한 kernel 함수의 가중치가 결정이 되므로 smoothing parameter로 작용하여 전체적인 회귀모형에 영향을 미치게 된다.

3. Bandwidth 추정

적정한 Kernel regression의 수행을 위한 bandwidth λ의 추정법으로 가장 많이 사용되는 것이 교차확인(Cross Validation)이다. 이는 회귀모형의 MSE(Mean Squared Error)를 최소화 할 수 있는 bandwidth를 설정하는 개념에서 출발하는 것이다.

3.1 Least Squares method

LS 방법은 자료와 추정된 값의 MSE를 최소로 하는 λ를 찾아내는 것으로 다음 식과 같이 정의된다.

관측값으로부터 추정된 f의 $x=x_j$에서의 추정치를 $f_\lambda(x_j)$라 하면 $LS(\lambda)$는 다음과 같다.

$$LS(\lambda)=\frac{1}{n}\sum_{j=1}^{n}(y_j-f_\lambda(x_j))^2 \tag{7}$$

여기서, y_i는 i번째 관측자료.

3.2 Least Squares Cross Validation

n개의 관측치중 j 번째를 제외한 n-1 개의 관측치로 추정된 f의 $x=x_j$에서의 추정치를 $f_{\lambda(j)}$라 하면 $CV(\lambda)$는 다음 식 8과 같이 정의된다.

$$CV(\lambda)=\frac{1}{n}\sum_{j=1}^{n}(y_j-f_{\lambda(j)}(x_j))^2 \tag{8}$$

소거정리에 의하여 식 8을 변형하면, 다음의 식 9와 같다.

$$CV(\lambda)=\frac{1}{n}\sum_{j=1}^{n}\left(\frac{y_j-f_\lambda(x_j)}{1-h_{jj}}\right)^2 \tag{9}$$

λ가 어떤 값으로 주어졌을 때 $f(x)$의 추정치 $\hat{f}_\lambda(x)$라고 하면, 다음과 같은 식을 사용할 수 있다.

$$\hat{f}_\lambda(x)=H_\lambda y \tag{10}$$

여기서, H_λ ; hat matrix , h_{jj}는 H_λ의 {j,j}

즉, Cross-Validation 방법은 추정치 n-1개의 관측치만 써보고, 1개의 관측값으로 확인(validation)을 해보는 기법이다. $CV(\lambda)$값을 최소로 하는 λ값이 최적의 bandwidth가 되며, Least squares cross-validation은 MSE의 bias를 적게 조절하여 주는 방법이다.

3.3 General Cross Validation

GCV는 LSCV의 개념과 같은 방식이지만, h_{jj} 대신에 $tr(H_\lambda)/n$ 을 사용한 것이다(Wand 등, 1995). $GCV(\lambda)$는 다음 식 11과 같이 정리 할 수 있다.

$$GCV(\lambda) = \frac{1}{n} \sum_{j=1}^{n} \left(\frac{y_j - f_\lambda(x_j)}{1 - tr(H_\lambda)/n} \right)^2 \tag{11}$$

or

$$GCV(\lambda) = \frac{\frac{1}{n}\left(\sum_{j=1}^{n} (y_j - f_\lambda(x_j))^2 \right)}{\left(\frac{1}{n}\ tr[I - H_\lambda]\right)^2} \tag{12}$$

$$= n \times \frac{\left(\sum_{j=1}^{n} (y_j - f_\lambda(x_j))^2 \right)}{(EDF)^2} \tag{13}$$

· 여기서, $tr(H_\lambda)$ 는 $\sum_{j=1}^{n} h_{jj}$, $\left(\sum_{j=1}^{n} (y_j - f_\lambda(x_j))^2 \right.$; residual sum of squares,

$EDF = tr[I - H_\lambda]$; equivalent degrees of freedom.

LSCV와 개념적으로 동일선상에 있는 GCV의 경우는 LSCV 보다 전체적으로 bias를 적게 만드는 특징을 지니게 되며, noise가 큰 자료의 경우 bandwidth의 결정에 있어서 LSCV보다 적합한 것으로 알려져 있다(Eubank, 1988).

4. 적용 및 결과

4.1 적용

Kernel regression에 있어 bandwidth 결정방법에 의한 회귀결과를 비교하기 위하여, 임의의 함수를 설정하여 자료군을 생성하고, 생성된 자료에 normal 분포로부터 무작위 추출한 잡음(noise)을 첨가한 뒤, Kernel regression을 실시하였다.

사용된 함수는 다음 표 1과 같다.

표 1. 적용함수

Data	True Function	Noise	Sample size
1	$f(t)=2 \quad 0.5<t\le1$ $f(t)=1 \quad 0\le t\le0.5$	N(0,0.04)	100
2	$f(t)=2 \quad 0.5<t\le1$ $f(t)=1 \quad 0\le t\le0.5$	N(0,1)	100
3	$f(t)=e^{-t}\sin(2\pi t) \quad 0\le t\le1$	N(0,0.01)	100
4	$f(t)=e^{-t}\sin(2\pi t) \quad 0\le t\le1$	N(0,0.25)	100

4.2 결과

Kernel function은 Quadratic kernel을 사용하였으며, 회귀모형을 구한 결과는 다음 그림 1 ~ 그림 8과 같다. 그림에서 얇은 실선은 원함수를 나타내며 굵은 실선은 Kernel regression에 의한 회귀모형을 나타낸다.

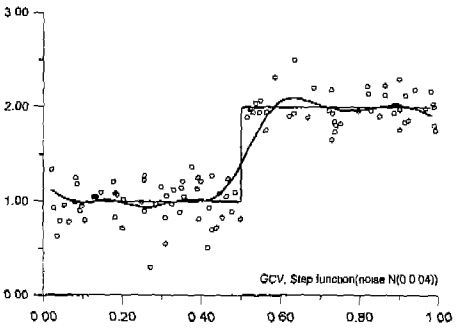

그림 1. GCV를 선택한
Kernel regression 결과
(Step function, noise(N(0,0.04))

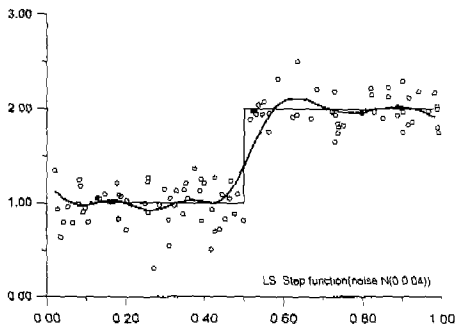

그림 2. LS를 선택한
Kernel regression 결과
(Step function, noise(N(0,0.04))

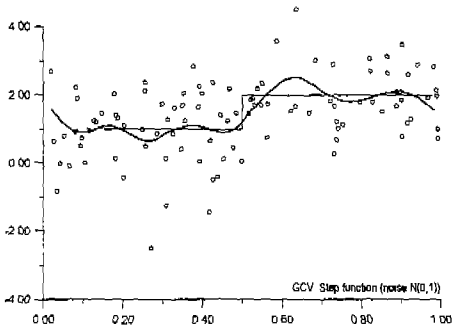

그림 3. GCV를 선택한
Kernel regression 결과
(Step function, noise(N(0,1))

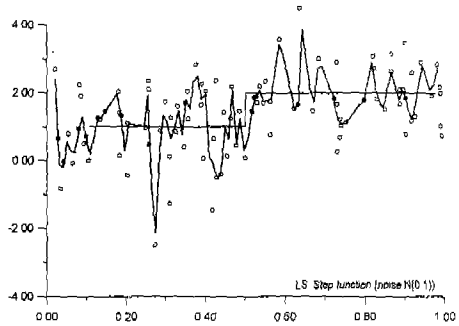

그림 4. LS를 선택한
Kernel regression 결과
(Step function, noise(N(0,1))

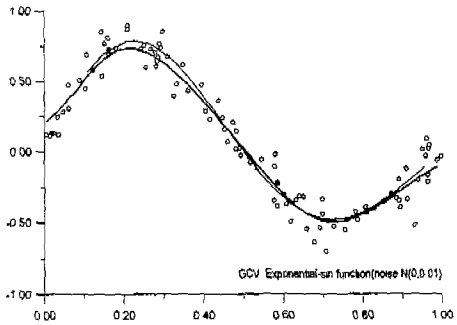

그림 5. GCV를 선택한
Kernel regression 결과
Exponential-sin function(noise(0,0.1))

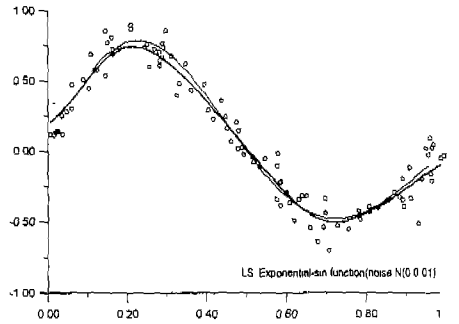

그림 6. LS를 선택한
Kernel regression 결과
Exponential-sin function(noise(0,0.1))

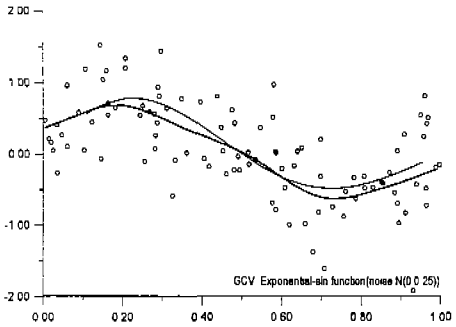

그림 7. GCV를 선택한
Kernel regression 결과
Exponential-sin function(noise(0,0.25))

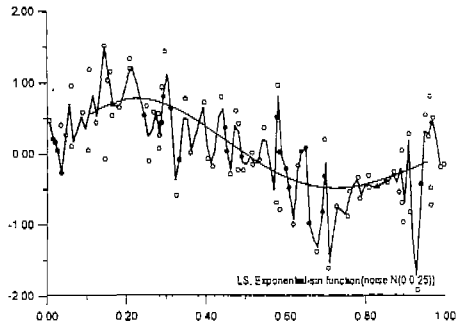

그림 8. LS를 선택한
Kernel regression 결과
Exponential-sin function(noise(0,0.25))

5. 결론

위 결과에서 볼 수 있듯이 Kernel regression은 자료의 분포 및 경향성을 반영한 적절한 회귀모형을 제시한다고 할 수 있다. GCV와 LS에 의한 Kernel regression 결과는 GCV가 LS에 비하여 전반적으로 bias와 MSE가 작고, 특히 그림 4, 8와 같이 잡음(noise)이 크면, LS는 bandwidth h값을 매우 작게 설정하게 되어 부적합한 회귀모형을 제시하게 되지만, GCV의 경우 그림 3, 7과 같이 적절한 모형을 회귀해내고 있다.

즉, LS의 경우 잡음(noise)이 큰 자료의 처리를 하게 될 경우, 오차의 bias가 커지게 되어 적정한 bandwidth를 선정하지 못하나, GCV의 경우는 Cross-Validation을 통해 오차의 bias를 줄임으로써 적합한 bandwidth를 선정하게된다.

따라서, 적절한 Bandwidth 선정방법인 GCV를 이용한 Kernel regression은 수문 기초자료의 효율적인 자료 처리 방법으로서 수자원분야에 광범위하게 이용될 수 있을 것이다.

6. 참고문헌

강근석, 김충락(1999) 회귀분석, 교우사, pp. 370~390.

문영일, 조성진, 김동권(2000) 수문학적 응용을 위한 비매개변수적 회귀모형 산정, 2000 학술발표회 논문집(Ⅲ), 대한토목학회, pp. 111~114.

조성진, 문영일, 권현한(1999) 비매개변수적 다항식을 이용한 수위-유량 관계곡선, 1999 학술발표회 논문집(Ⅲ), 대한토목학회, pp. 29~32.

Eubank, R.L. (1988) *Spline Smoothing and Nonparametric Regression*. DEKKER, New York.

Scott, David W. (1992) *Multivariate Density Estimation*. JOHN WILEY & SONS, New York.

Wand, M.P. and Jones, M.C. (1995) *Kernel Smoothing*. CHAPMAN & HALL, New York.