Journal of Statistical
Theory & Methods
1996, Vol. 7, pp 257~262

일반화최대우도함수에 의해 추정된 평활모수에 대한 진단

정원태 • 이인석 • 정혜정

요약 본 논문은 스플라인 회귀모형에서 평활모수를 추정할 때 사전 작업으로 영향력 진단을 하는 문제를 다룬다. 평활모수의 추정방법으로 일반화최대우도함수법을 사용할 때, 얻어지는 추정치에 영향을 주는 관측치를 진단하는 측도를 제안하고, 찾아낸 영향력 관측치를 수정하여 올바른 평활모수 추정치를 찾는 방법을 소개한다.

주제어 : 스플라인모형, 국소적진단, 평활모수, 가법적섭동

1. 서론

비모수회귀자료의 분석에 있어서 스플라인 함수가 많이 이용되고 있다. 이 스플라인 함수를 이용한 분석이 잘 이루어지기 위해서는 스플라인 함수의 차수(degree)와 평활(smoothing)정도를 조절해주는 평활모수(smoothing parameter)의 결정이 매우 중요하다고 할 수 있다. 그러나 차수는 3차에서 대부분의 자료를 적합화 시킬 수 있다는 사실이 80년대 초반의 논문들에서 발표가 되었으며, 이후 Whaba(1985), Gasser(1991) 등의 논문들은 주로 평활모수의 추론에 관심을 가지고 있다. 평활모수에 대한 추정법은 교차타당성법(Cross-Validation), 일반화교차타당성법(generalized Cross-Validation) 그리고 일반화최대우도추정법(generalized maximum likelihood estimation) 등이 주로 사용되고 있다.

추론을 포함하여 대부분의 경우에서 자료의 적합성을 평가할 때, 사전작업으로 회귀진단을 많이 하고 있으며, 이 분야 또한 많이 발전하여 왔다. 회귀진단에는 크게 두가지 방법을 통하여 이루어지고 있다고 할 수 있다. 첫째는 주어진 자료에서 하나씩 또는 여러개를 빼가며 추론을 하여 모든 자료를 모두 사용하여 얻어진 추론의 결과와 비교를 통하여 이상치나 영향치를 찾아내는 방법이며, 둘째는 모든 관측치에 동시에 약간의 제약을 가하여 추론의 결과를 살펴보는 방법인데, 이 방법은 Cook(1986)에 의하여 제안된 국소적진단법이다.

지금까지 발표된 논문들 중 스플라인 회귀모형에서 회귀함수의 추론에 대해서는 회귀진단을 다루었으며 평활모수의 추론에 대한 진단은 소개된 3가지의 방법중 일반화교차타당성법에 대한 Lee and Jung(1995)의 논문 외에는 거의 소개되지 않았다.
따라서 우리는 이 논문을 통하여 일반화최대우도추정법에 의한 평활모수를 추정할 때의 사전작업으로 회귀진단을 하여보고자 한다.

2절에서는 스플라인회귀모형과 이 회귀모형에 따른 일반화최대우도함수식을 소개하고, 또한 Cook(1986)에 의하여 소개된 국소적(local)진단법을 평활모수의 추정에 사용할 일반화최대우

정원태는 경북대학교 통계학과(702-701 대구광역시 북구 산격동 1370번지)의 강사이다. 이인석은 경북대학교 통계학과의 교수이다. 정혜정은 평택대학교 전산통계학과 (405-701 경기도 평택시 용이동 111번지)전임강사이다.

도추정법에 적용하여 본다. 그리고 3절에서는 관측치에 가법적섭동(additive perturbation) 상황을 고려하여 일반화최대우도함수를 정의하고 이 섭동 상황에서 영향을 크게 주는 관측치를 판단할 수 있는 진단측도를 제안한다.

4절에서는 제안된 진단측도를 사용하는 방법과 찾아낸 영향치들을 수정하여 올바른 평활모수를 찾는 방법에 대해서 소개한다.

2. 회귀모형과 국소적 진단법

다음과 같은 상황을 생각해보자. 관측치 $y_1,\cdots,y_n$은 편의상 $a \le t_1 <\cdots< t_n \le b$의 관계를 만족하는 독립변수 $t_1,\cdots,t_n$에 대응되며, 회귀모형은

$$y_j = \mu(t_j) + \varepsilon_j \tag{1}$$

와 같이 주어진다. 여기서 μ는 어떤 평활회귀함수이며, 오차 ε_j들은 평균 0, 분산 σ^2을 가지면서 서로 무상관이다. 평활함수란 $W_2^m[a,b]$에 속하는 함수이며, $W_2^m[a,b]$는 어떤 고정된 m에 대하여 $m-1$번째까지의 절대적으로 연속인 도함수를 가지고, m번째 도함수의 자승이 주어진 구간에서 적분가능한 집합을 말한다.

위의 가정하에서 회귀함수의 추정량 $g \in W_2^m[a,b]$는 일반적으로 아래 식을 최소화 하는 방법으로 얻어진다.(Eubank(1988))

$$\frac{1}{n}\sum_{j=1}^{n}\{y_j - g(t_j)\}^2 + \lambda\int_a^b\{g^{(m)}(t)\}^2\,dt,\ \lambda > 0. \tag{2}$$

이 때 $n \ge m$인 경우, (2)를 만족하는 추정량을 평활스플라인(smoothing spline)이라 한다.

식(2)로부터 자료의 평활 정도를 결정하는 모수λ의 결정은 중요한 문제이다. 이 평활모수를 추정하는 방법으로 일반화최대우도법이 많이 사용되고 있는데, 이 일반화최대우도법은 Wahba(1985)와 Barry(1995) 등이 사용하였으며, 식(2)에 따른 일반화최대우도함수의 수식은 아래와 같다.

$$GML(\lambda) = \frac{\underset{\sim}{y}^T(I - H_\lambda)\underset{\sim}{y}}{[\det(I - H_\lambda)]^{1/(n-1)}} \tag{3}$$

여기서 H_λ는 선형회귀모형에서의 "Hat"행렬로 3절에서 다시 소개되며, det(.)는 행렬식을 의미한다.

식(3)을 최소로하는 평활모수를 찾는 방법이 일반화최대우도함수법이며 이후로는 $\hat{\lambda}_{GML}$을 이 방법에 의해서 추정된 평활모수의 추정치로 표시하기로 한다.

이제 Cook(1986)의 국소적 영향력 진단법을 일반화최대우도함수에 의한 추정법에 적용하여, 일반화최대우도함수법에 의해 얻어지는 추정치 $\hat{\lambda}_{GML}$에 대한 진단을 해보자.

먼저 관심있는 모수 β 와 n개의 관측자료의 집합 $\underset{\sim}{y}$를 가지는 통계모형을 고려해보자. 그리고 모수 β 의 어떤 함수를 $L(\beta|\underset{\sim}{y})$라 두자. 물론 이 함수로 우도함수를 사용할 수도 있다. 우선 주어진 모형에 도입되는 수학적인 양으로 섭동(perturbation)을 $\underset{\sim}{\omega}$로 두자. 이 $\underset{\sim}{\omega}$는 모수도 아니고 자료도 아니다. 그리고 $\hat{\beta}(\underset{\sim}{\omega})$를 $\underset{\sim}{\omega}$ 섭동이 있는 경우에 얻어지는 모수 β 의 추정치로 두고, $\hat{\beta}(\underset{\sim}{\omega_0})$ 는 섭동이 없는 경우에 얻어지는 추정치라 하자. 여기서 영향력 평가는 $\hat{\beta}(\underset{\sim}{\omega})$ 와

$\hat{\beta}(\underline{\omega}_0)$ 의 비교를 통하여 얻어진다. 즉, 무(null) 섭동인 $\underline{\omega}_0$ 로부터 아주 조금의, 이후로는 국소적(local), 위반(departure)인 섭동 $\underline{\omega}$ 를 사용할 때 모수의 추정에 영향을 가장 많이 미치는 관측치를 찾아내는 것이다. 이 영향력 정도를 평가하는 도구로써 Cook(1986)과 Lawrance(1991)는 아래와 같이 주어지는 $t_{\max}$ 를 사용하였다.

$$t_{\max} \propto \frac{\partial \hat{\beta}(\underline{\omega})}{\partial \underline{\omega}^T}|_{\underline{\omega}_0}$$

따라서 우리는 앞에서 주어진 스플라인 회귀모형에서 $\hat{\lambda}(\underline{\omega})$ 를 섭동이 있는 경우에 추정된 평활모수라 두고 $\hat{\lambda}(\underline{\omega}_0)$ 를 무 섭동 상태에서 얻어진 추정치라고 한다면, 우리는 쉽게

$$t_{\max} \propto \frac{\partial \hat{\lambda}(\underline{\omega})}{\partial \underline{\omega}^T}|_{\underline{\omega}_0} \tag{4}$$

를 이용하여 얻어진 진단측도를 구할 수 있다.
이 $t_{\max}$ 는 어떤 자료가 국소적 섭동상태일 때 얼마나 추론에 변화를 일으키는 가에 대한 측도로, 이 값의 절대값이 큰 관측치를 이 추론에 영향을 주는 관측치로 판별하는 것이다. 이 $t_{\max}$ 를 Cook(1986) 등 여러 저자들은 최대기울기방향(direction of maximum slope)이라 부른다.

우리가 여기서 평활모수 λ의 추정법으로 일반화최대우도법을 사용한다면 $\hat{\lambda}(\underline{\omega})$는

$$\frac{\partial GML(\lambda,\underline{\omega})}{\partial \lambda}|_{\lambda=\hat{\lambda}(\omega_0)} = 0 \tag{5}$$

의 해로 얻어진다.
여기서 $GML(\lambda,\underline{\omega})$은 섭동이 포함된 일반화최대우도함수이며

$$GML(\lambda,\underline{\omega}) = \frac{\underline{y}_\omega^T (I - H_\lambda)\underline{y}_\omega}{[\det(I - H_\lambda)]^{\frac{1}{n-1}}}$$

와 같이 정의 된다. 여기서 $\underline{y}_\omega$ 는 섭동이 포함된 관측치 벡터이다. 따라서 이 식을 이용하여 우리는 다음과 같은 유용한 정리를 얻을 수 있다.

정리 1. $GML(\lambda,\underline{\omega})$를 위의 식에서 정의된 함수라 두면, 최대기울기방향은 다음과 같이 얻어진다.

$$t_{\max} \propto \frac{\partial^2 GML(\lambda,\underline{\omega})}{\partial \underline{\omega}^T \partial \lambda}|_{\underline{\omega}_0,\hat{\lambda}}.$$

증명. 함축함수(implicit function) 정리를 이용하여 (5)의 양변을 미분하면

$$\frac{\partial \hat{\lambda}(\underline{\omega})}{\partial \underline{\omega}^T} = (\frac{\partial^2 GML}{\partial \lambda^2})^{-1}(\frac{\partial^2 GML}{\partial \lambda \partial \underline{\omega}^T})$$

이고, 오른쪽의 첫항은 스칼라로 무시할 수 있다. 따라서 식(4)로 부터 위의 결과를 얻을 수 있다.

3. 진단측도

관측치에 가법적으로 국소적 섭동이 포함된 형태로, $\underline{y}_{\omega} = \underline{y} + \underline{\omega}$ 와 같이 표현되는 관측치의 벡터를 생각해보자. 여기서 $\underline{y}$는 섭동이 없는 경우에 얻을 수 있는 관측치이고, $\underline{y}_{\omega}$ 는 $\underline{\omega}$ 의해서 섭동이 생긴 경우에 관측되는 실제 관측치 이다. 여기서 $\underline{\omega} = (\omega_1, \cdots, \omega_n)$이고, 섭동이 없는 경우, $\underline{\omega}_0 = (0, \ldots, 0)$이다. 위에서와 같은 방법으로 섭동이 관측치에 주어지는 경우를 가법적(additive) 섭동 상황이라 한다. 가법적모순의 경우 외에도 여러개의 섭동 상황은 Emerson(1984) 등과 Lawrance(1991)에 의해서 소개되고 있다.

이 섭동 상황하에서 일반화최대우도법에 의해 추정된 평활 모수 $\hat{\lambda}_{GML}$ 는

$$GML(\lambda, \underline{\omega}) = \frac{(\underline{y} + \underline{\omega})^T (I - H_\lambda)(\underline{y} + \underline{\omega})}{[\det(I - H_\lambda)]^{\frac{1}{(n-1)}}}$$

을 최소화 해주는 λ를 구하여 주면 된다.

따라서 정리1에 의해 가법적 섭동상황에서 추정치에 영향을 많이 주는 관측치를 판별하는 최대 기울기방향은 다음과 같이 구해진다.

아래의 계산을 위하여 'Hat'행렬의 구조를 살펴보자. 먼저 s를 추정을 위해서 사용하는 스플라인 함수라 두고, 다음과 같은 식을 만들어보자.

$$Q(s,t) = [(m-1)!]^{-2} \int_a^b (s-u)_+^{m-1} (t-u)_+^{m-1} \, du, \quad a \le s, t \le b.$$

여기서

$$(x)_+ = \begin{cases} x, & x > 0 \\ 0, & elsewhere \end{cases}$$

그리고 이 수식을 이용하여

$$Q_n = \{Q(t_i, t_j)\}_{i,j=1,\cdots,n}$$

을 만들면, 이 행렬은 양정치 행렬이 된다.

Wahba(1978)는 $U^T U = I$가 되는 U를 이용하여 회귀함수의 추정량으로 $\underline{\mu}_\lambda = H_\lambda \underline{y}$ 을 사용하였다. 여기서 $H(\lambda) = I - n\lambda U(U^T Q_n U + n\lambda U^T U)^{-1} U^T$ 이다. 이를 이용하여 아래와 같이 가법적 섭동이 존재하는 경우를 진단할 수 있는 측도를 구할 수 있다.

$$\begin{aligned} t_{\max}(\underline{y} + \underline{\omega}) &\propto \frac{\partial^2 GML(\underline{\omega}, \lambda)}{\partial \underline{\omega}^T \partial \lambda}, \\ &\propto c_2 M_\lambda y - c_2 c_3 M_\lambda^3 y \\ &\propto c_2 M_\lambda (I - c_3 M_\lambda^2) y \end{aligned}$$

여기서

$$M_\lambda = (I - H_\lambda), \; c_2 = \frac{2}{\lambda} |M_\lambda|^{-1}, \; c_3 = \frac{1}{n-1} |M_\lambda^{-1}| |M_\lambda|^{\frac{2-n}{n-1}}$$

이며, λ 값으로는 처음 주어진 관측값들로 부터 구한 초기값을 사용한다. $t_{\max}(\underline{y} + \underline{\omega})$는 관측값의 수 많큼의 차수를 가지는 벡터로서 각각의 관측치들이 국소적 섭동에 의하여 영향을 주는 정

도를 나타낸다. 즉 각각의 진단측도값의 절대치들중 3번째와 10번째가 가장 큰 값이었다면 3번째와 10번째의 관측치가 추론에 영향을 가장 많이 주는 것으로 평가된다.

4. 제언 및 결론

3절에서 제안된 진단측도는 스플라인 회귀모형에서 평활모수의 추정법으로 일반화최대우도법을 사용하는 경우에 사용할 수 있는 진단측도이다. 즉 일반화최대우도법에 의해 추정되는 평활모수가 올바르게 추정되기 위한 사전작업으로 사용할 수 있는 측도이다. 이 측도에 의해서 찾아지는 관측치들은 섭동에 의한 반응 정도, 즉 $t_{\max}(\underset{\sim}{y}+\underset{\sim}{\omega})$의 해당 값을 통하여 양으로 크게 나오는 케이스는 그 해당 관측값에 ω_i 만큼 빼주고, 음으로 크게 나오는 케이스는 그 해당값에 ω_i 만큼 더해주어 추정치에 영향을 크게 미치지 못하게 한 다음 평활모수를 추정해준다. 반응값이 크게 나오는 관측치를 찾는 기준은, 즉 진단측도에서 얻어진 값 중에 그 정도가 크다고 할 수 있는 것을 선택하는데, 그 기준으로 상자-그림을 이용하는 것도 하나의 방법이다.

참고자료

Barry, D. (1995), *A Bayesian Model for Growth Curve Analysis*, Biometrics 51, pp. 639-655.

Cook, R. D. (1986), *Assessment of Local Influence(with Disscussion)*, Journal of the Royal Statistical Society, B, Vol. 48, pp. 133-169.

Emerson, J. D., Hoaglin, D. C. and Kempthorne, P. J. (1984), *Leverage in Least Squares Additive-Plus-Multiplicative Fits for Two-Way Tables*, Journal of the American Statistical Association, Vol. 79, pp. 329-335.

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

Gasser, T., Kneip. A., and Kohier, W. (1991), *A Flexible and Fast Method for Automatic Smoothing*, Journals of the American Statistical Association, 86, pp. 643-652.

Lee, I. S. and Jung, W. T. (1995), *Diagnostic for Smoothing Parameter Estimate in Nonparametric Regression Model*, The Korean Communication in Statistics, V. 2, pp. 266-276.

Lawrance, A. J. (1991), *Directions in Robust Statistics and Diagnostics*, Splinger-Verlag, New York.

Wahba, G. (1978), *Improper Priors, Spline Smoothing, and the Problem of Guarding against Model Errors in Regression*, Journal of the Royal Statistical Society, B, 49, pp. 364-372.

Wahba, G. (1985), *A Comparision of GCV and GML for Choosing the Smoothing Parameter in the Genralized Spline Smoothing Problem*, Annals of Statistics, 13, pp. 1378-1402.

Diagnostics for Estimated Smoothing Parameter by Generalized Maximum Likelihood Function

Won Tae Jung · In Suk Lee · Hae Jeong Jeong

Abstract When we are estimate the smoothing parameter in spline regression model, we deal the diagnostic of influence observations as posteriori analysis. When we use Generalized Maximum Likelihood Function as the estimation method of smoothing parameter, we propose the diagnostic measure for influencial observations in the obtained estimate, and we introduce the finding method of the proper smoothing parameter estimate .

Keywords: spline model, local diagnostic, smoothing parameter, additive perturbation.

Won Tae Jung is a lecturer, Dept. of Statistics, Kyungpook University, Taegu, Korea 702-701.
Hae Jeong Jeong is a professor, Dept. of Computer Science & Statistics, Pyung Taek University, Kyunggido, Korea.