▩ 연구논문

Influence Diagnostic Measure for Spline Estimator

In-Suk Lee · Gyo-Young Cho · Won-Tae Jung
Dept. of Statistics, Kyungpook National University

Abstract

To access the quality of a fit to a set of data it is always useful to conduct a posteriori analysis involving the examination of residuals, detection of influential data values, etc. Smoothing splines are a type of nonparametric regression estimators for the diagnostic problem. And leverage value, Cook's distance, and DFFITS are used for detecting influential data. Since high leverage points will always have small residuals, the new diagnostic measures including of properties of leverage and residuals are needed. In this paper, we propose FVARATIO version as diagnostic measure in nonparametric regression. Also we consider the rough bound as analogy with linear regression case.

1. Introduction

Consider the regression model

$$y_i = \mu_\lambda(t_j) + \varepsilon_j \qquad j=1, \cdots, n \tag{1}$$

where $a \le t_1 < \cdots < t_n \le b$ and the ε_j are uncorrelated random errors with mean zero and variance σ_λ^2. In this paper we consider the problem of nonparametric estimation of the regression function μ_λ. Diagnostic methods for a spline estimator of regression function are proposed and their properties are investigated.

It will be assumed that μ_λ is smooth in the sense that, for positive integer m, μ_λ admits $m-1$ continuous derivatives and has a square integrable mth derivative. Under these restrictions a natural estimator is the function $\hat{\mu}_\lambda$ which minimizes

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

Several procedures are available for the estimation of smoothing parameter, λ, from data. Craven and Wahba(1979) considered the use of the generalized cross validation for this purpose. That is, the choice of λ is to minimize

$$GCV(\lambda) = n^{-1}\sum_{i=1}^{n} e_{\lambda}^{2}(t_i)/\{1-n^{-1}tr(H(\lambda)\}^{2},$$

where $e_{\lambda}(t_i)=y_i-\mu_{\lambda}(t_i)$, tr denotes the matrix trace, and $H(\lambda)$ is the $n\times n$ matrix which transforms the vector of responses, $\underline{y}$, to the vector of fitted values. In this paper attention will be restriced to the estimator $\hat{\mu}_{\hat{\lambda}}$ with $\hat{\lambda}$ estimated by GCV.

The adverse influence that outliers can have on smoothing spline estimates has been recognized by many authors. Their solution to such difficulties is the use of robust smoothing splines. In contrast, the objective of this paper is the development of techniques for detection of influential data, observations which significantly affect the fit.

In Section 2, we present three diagnostic measures of $\hat{\mu}_{\hat{\lambda}}$. In Section 3, a diagnostic measure for spline estimator of regression function is proposed. The use of this measure is illustrated in Section 4 through a numerical example.

2. Review of Influence Diagnostic Measures

The vector of residuals in the regression model (1),

$$\underline{e}_{\hat{\lambda}} = (e_{\hat{\lambda}}(t_1), \cdots, e_{\hat{\lambda}}(t_n))' = \underline{y}-\underline{\mu}_{\hat{\lambda}}$$

has variance-covariance matrix $Var(\underline{e}_{\hat{\lambda}})=\sigma^2(I-H(\hat{\lambda}))$. One might use standardized residuals $r_{\hat{\lambda}_i}(\sigma)=e_{\hat{\lambda}}(t_i)/\sigma(1-h_{ii}(\hat{\lambda}))^{1/2}$ to assess the quality of the fit to y_i. That is, we can find the outliers by using of these standardized residuals. However as noted by many authors an outlier needs not be influential. Thus we need other measures for detection of influential data. Now we introduce some known diagnostic measures for the estimator of regression function in spline smoothing.

2.1 Leverage Measure

Because smoothing spline is linear estimator, there is an $n\times n$ matrix $H(\hat{\lambda})=\{h_{ij}(\hat{\lambda})\}$ which transforms the vector of responses, $\underline{y}$, to the vector of fitted values $H(\hat{\lambda})$ is known as the hat matrix(Eubank(1984)) and its element, $h_{ij}(\hat{\lambda})$ determines

how much influence y_i has on the fit to y_i. By analogy with the linear regression setting, Eubank(1984) studied for diagonal elements of $H(\hat{\lambda})$ where it is shown, for example, that $0 \le h_{ii}(\hat{\lambda}) \le 1$. By analogy with the linear regression case we might say that an observation has leverage if its leverage value exceeds $3tr\{H(\hat{\lambda})\}/n$ or $\{2trH(\hat{\lambda})+1\}/n$.

2.2 Cook's Distance Measure

Eubank(1988) took the Cook's measure as $D_{\hat{\lambda}_i} = r_{\hat{\lambda}_i} h_{ii}(\hat{\lambda})/(1-h_{ii}(\hat{\lambda}))trH(\hat{\lambda})$ in spline smoothing. And one might deem an observation influential if its corresponding Cook's distance exceeds the lower 10% point of the F-distribution with approximate numerator and denomonator degrees of freedom $trH(\hat{\lambda})$ and $tr(I-H(\hat{\lambda}))$, respectively.

2.3 DFFITS Measure

The diagnostic measure, $DFFITS$ in Eubank(1985) and Silverman(1985) is proposed by

$$DFFITS_i = r_{\hat{\lambda}_i}(\hat{\sigma}_{\hat{\lambda}}) \cdot \{h_{ii}(\hat{\lambda})/(1-h_{ii}(\hat{\lambda}))\}^{1/2}.$$

Here he considered the use of rough bound by $2\{tr(H(\hat{\lambda})/tr(I-H(\hat{\lambda}))\}^{1/2}$.

3. The Proposed Influence Diagnostic Measure

We now propose the new diagnostic measure, such as, $FVARATIO$ in spline smoothing. Let $\sigma^2_{\hat{\lambda}}$ and $\sigma^2_{\hat{\lambda}(i)}$ denote variances obtained from the entire data and from the data when the ith case has been deleted, respectively. Under the model (1), an unbiased estimator $\hat{\sigma}^2_{\hat{\lambda}}$ of $\sigma^2_{\hat{\lambda}}$ is provided by

$$\hat{\sigma}^2_{\hat{\lambda}} = \sum_{i=1}^{n} e_{\hat{\lambda}}(t_i)^2/tr(I-H(\hat{\lambda})). \tag{3}$$

The next theorem is useful for calculation of $\hat{\sigma}^2_{\hat{\lambda}(i)}$.

Theorem 3.1 (Eubank and Gunst(1986); Deletion Theorem) For fixed λ and z, let $\hat{\beta}(\lambda, z)$ denote the minimizer of

$$n^{-1}\{\sum_{j\neq i}(y_i - x_j'\beta)^2 + (z - x_i'\beta)^2\} + \lambda\beta' G\beta.$$

Then $\hat{\beta}_{(i)}(\lambda) = \hat{\beta}(\lambda, x_i'\hat{\beta}_{(i)}(\lambda))$ and $x_i'\hat{\beta}_{(i)} = y_i - e_i(\lambda)/(1-h_{ii}(\lambda))$.

It follows from Deletion Theorem that $\hat{\sigma}^2_{\lambda(i)}$ is the estimator of $\hat{\sigma}^2_{\lambda}$ computed from the data when the observation (t_i, y_i) has been deleted. That is,

$$\hat{\sigma}^2_{\lambda(i)} = \sum_{j\neq i} \{e_\lambda(t_j) + h_{ji}(\hat{\lambda})e_\lambda(t_i)/(1-h_{ii}(\hat{\lambda}))\}^2/(n-1-tr[H_{(i)}(\hat{\lambda})]), \tag{4}$$

where

$$tr[H_{(i)}(\hat{\lambda})] = \sum_{j\neq i} \{h_{jj}(\hat{\lambda}) + h_{ij}(\hat{\lambda})^2/(1-h_{ii}(\hat{\lambda}))\}. \tag{5}$$

Using estimators in (3) and (4), we can obtain

$$Var(\hat{y}(t_i)) = h_{ii}(\hat{\lambda})\sigma^2_\lambda,$$

and

$$Var(\hat{y}_{(i)}(t_i)) = h_{ii}(\hat{\lambda})\sigma^2_\lambda/(1-h_{ii}(\hat{\lambda})). \tag{6}$$

In the case that σ^2_λ is unknown it may be replaced by either $\hat{\sigma}^2_\lambda$ or $\hat{\sigma}^2_{\lambda(i)}$ in (6). By analogy with the linear regression case, we define

$$\begin{aligned} FVARATIO_i^* &= \widehat{Var}(\hat{y}(t_i))/\widehat{Var}(\hat{y}_{(i)}(t_{(i)})), \\ &= \frac{e^2_\lambda(t_i)}{r^2_\lambda \sum_{j\neq i}\{e_\lambda(t_j) + h_{ji}(\hat{\lambda})e_\lambda(t_i)/(1-h_{ii}(\hat{\lambda}))\}^2/\tau(i)}, \\ &= \frac{(1-h_{ii}(\hat{\lambda}))\sigma^2_\lambda}{\sum_{j\neq i}\{(1-h_{ii}(\hat{\lambda}))e_\lambda(t_j) + h_{ji}(\hat{\lambda})e_\lambda(t_i)\}^2/\tau(i)}, \end{aligned} \tag{7}$$

where $\tau(i) = (n-1-tr[H_{(i)}(\hat{\lambda})])$.

Let p be the dimension of regression coefficients in parametric regression case. Then by replacing p with trace of hat matrix, as procedures of Eubank(1985), we can propose the rough bound as following

$$FVARATIO_i^* \leq 1 - 3/\tau(i)$$

or

$$FVARATIO_i^* \geq 1+2tr(H_{(i)}(\hat{\lambda})+1)/\tau(i), \tag{8}$$

where $\tau(i)$ is given in (7).

4. Example and Summary

As an illustration of the use of the diagnostics discussed in Section 3 for analysis of data by smoothing splines we consider the German hyperinflation data. The data consist of values for the logarithm of the money supply as a function of the logarithn for the premium, or discount, on a forward contract for foreign exchange during the German hyperinflation(see Wecker and Ansley, 1983). A cubic smoothing spline was fitted to this data with the smoothing parameter value $\hat{\lambda}$ selected by GCV.

〈 Table 1 〉 Result of Influential Diagnostic

	Influence Measure	Influence Observation
Smoothing Parameter Estimate is 2.45×10^{-4}	Cook's Distance	19 24 28 29 31
	DFFITS	19 24 28 29 31
	LEVERAGE	18 19 31
	*FVARATIO**	18 19 24 30 31

In 〈Table 1〉, observations, 18, 19, and 31 are high leverage data points and observations 24, 28, 29, and 31 are seen to be influential because they have large standarized residuals, that is, Cook's and DFFITS measures are the functions of standarized residual. Since high leverage points will always have small residuals, we need the measures which have above three diagnostic properties. In results of 〈Table 1〉, *FVARATIO** is seen such measure. Their influence was clearly revealed by examination of *FVARATIO**.

References

[1] Craven, P. and Wahba, G. (1979), "Smoothing Noisy Data with Spline Functions; Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation," *Numerical Mathematics*, Vol. 31, pp. 377-403.

[2] Eubank, R. L. (1984), "The Hat Matrix for Smoothing Splines," *Statistics and Probability Letters*, Vol. 2, pp. 9-14.

[3] Eubank, R. L. (1985), "Diagnostics for Smoothing Splines," *Journal of the Royal Statistical Society*, B, Vol. 47, pp. 332-341.

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

[5] Eubank, R. L. and Gunst, R. F. (1986), "Diagnostics for Penalized Least Squares Estimators," *Statistics and Probability Letters*, Vol. 4, pp. 265-272.

[6] Silverman, B. W. (1985), "Some Aspects of the Spline Smoothing Approach to Non-parametric Regression Curve Fitting," *Journal of the Royal Statistical Society*, B, Vol. 47, pp. 1-52.

[7] Wecker, W. P. and Ansley, C. F. (1983), "The Signal Extraction Approach to Nonlinear Regression and Spline Smoothing," *Journal of the American Staistical Association*, Vol. 78, pp. 81-89.