Journal of Statistical
Theory & Mothods
1995, Vol. 6, pp 63~71

Diagnostic In Spline Regression Model With Heteroscedasticity

In-Suk Lee[1], Won-Tae Jung[1] and Hye-Jeong Jeong[2]

Abstract We have consider the study of local influence for smoothing parameter estimates in spline regression model with heteroscedasticity. Practically, generalized cross-validation does not work well in the presence of heteroscedasticity. Thus we have proposed the local influence measure for generalized cross-validation estimates when errors are heteroscedastic. And we have examined effects of diagnostic by above measures through Hyperinflation data.

Keywords : generalized cross validation, spline smoothing, local influence, smoothing parameter, heteroscedasticity, hat matrix.

1. Introduction

A standard assumption in both parametric and nonparametric regression is homogeneity of the error variances. Violation of this assumption can have adverse consequences for efficiency of estimators. Smoothing splines are a type of nonparametric regression for estimators which the diagnostic problem has received a good bit of attention. Whereas current diagnostic methods for smoothing spline are mostly of the case-deletion variety,. including parallels of residuals and Cook's distance measure. Eubank(1988), Silverman(1985), and Eubank and Thomas(1993) have treated diagnostic methods for smoothing splines with examples.

Assume that responses $y_1, \cdots, y_n$ are observed which follow the model

$$y_i = \mu(t_{in}) + \varepsilon_i, \qquad i = 1, \cdots, n,$$

where $0 \le t_{1n} < \cdots < t_{nn} \le 1$ and the errors ε_j are independent random variables with mean 0 and common variance σ^2, μ is a some smooth regression function.

---

[1] Dept. Statistics, Kyungpook National University.
[2] Dept. Computer Science and Statistics, Pierson University.

Also we are assumed

$$\mu \in W_2^m = \{g | g^{(i)} \ \ is \ \ absol. \ \ conti. \ \ and \ \ \int g^{(m)}(t)dt < \infty\}$$

for some integer $m \geq 1$.

For diagnostic and estimation of regression function by smoothing splines, the decision of the measure of smoothness is required. This measure is called the smoothing parameter. Therefore the selecting of the smoothing parameter is a crucial part of the fitting process. In particular, automatic procedures for selecting such tuning constants based on the data such as cross-validation CV), generalized cross-validation(GCV), etc., are often preferred.

And we use mainly estimator $\hat{\lambda}$ based on the assumptions of above is the minimizer over $g \in W_2^m$ of

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

In particular Li(1985), Hall and Titterington(1987) have employed generalized cross-validation to select *smoothing parameter* in problems of estimation. And the GCV choice $\hat{\lambda}$ which minimizes

$$GCV(\lambda) = \frac{\underline{e}_\lambda^T \underline{e}_\lambda}{\{tr(I - H(\lambda))\}^2} \tag{2}$$

where $H(\lambda)$ is the hat matrix that transforms the data vector $\underline{y}$ into the vector of smoothing spline fitted values, and $\underline{e}_\lambda = (I - H(\lambda))\underline{y}$ is the vector of residuals.

If the errors are heteroscedastic, then expressions(1) and (2) are no longer the correct estimation. The sum of squared error term in criterion (1) should include weight that are reciprocals of the variances of the ε_i. Practically, Hall and Titterington(1987) showed that GCV estimators of smoothing parameter can be slightly underestimated in both linear ridge regression and nonparametric regression. And Eubank and Thomas(1993) suggest caution in the routine use of GCV for smoothing parameter selection in the presence of heteroscedasticity.

Thus, we need to use the proper smoothing parameter to produce satisfactory results. That is, diagnostics for estimates of smoothing parameter are needed when the errors are heteroscedastic. Hence, the objective of this paper is to present diagnostics for influence on an important aspect of a fitted smoothing parameter by GCV under heteroscedasticity.

In Section 2, we introduce the local influence procedure for parameter estimates

in parametric regression and derive the local influence procedure for smoothing parameter estimates in nonparametric regression.

In Section 3, we obtain the diagnostic measure when the error are heteroscedastic.

In Section 4, we examine the effects of diagnostic through example.

2. Local Influence Procedure

Now, by the local-influence method of Cook(1986) and Lawrance(1991), we drive diagnostics which identify observation that have a disproportionately large impact on the determination of the GCV estimator $\hat{\lambda}$. Let Ω be some open set of allowable perturbations.

Suppose that the perturbation is $\underline{\omega}$ and the null perturbation is $\underline{\omega}_0$. In order to find direction of large local change, our first step is to approximate the actual surface with its tangent plane at $\lambda(\underline{\omega}_0)$ and find the direction of maximum slope $d_{\max}$ on this tangent plane.

It is easy to show that the direction of maximum slope is

$$d_{\max} \approx \frac{\partial \hat{\lambda}(\underline{\omega})}{\partial \underline{\omega}^T}$$

evaluated at $\underline{\omega}_0$. The direction vector $d_{\max}$ tells us how to perturb the data and the model to produce the greatest local change. Thus itself is the influence diagnostic measure, and the largest absolute components of $d_{\max}$ identify locally influential cases. Hence we have the following theorem .

Theorem 1. Let $GCV(\lambda, \underline{w})$ is the perturbed generalized cross validation function by $\underline{w}$. Then the direction of maximum slope is given by

$$d_{\max} \approx -\frac{\partial^2 GCV(\lambda, \underline{w})}{\partial \underline{w}^T \partial \lambda}$$

evaluated at $\hat{\lambda}$ and $\underline{w}_0$.

3. Diagnostic Measure

The criterion (1) for estimation is appropriate if each observation has the same variance. If this is not the case, the criterion should be modified by weighting observations inversely proportional to their standard deviations.

Hence we can propose a general criterion as

$$\min_{g \in W_2^m}[\frac{1}{n}\sum_{1}^{n}\{(1+w_j)(y_j - g(t_j)\}^2 + \lambda\int_a^b \{g^{(m)}(t)\}^2 dt] \tag{3}$$

with weight w_j such that $1+w_j = \{Var(y_j)\}^{-1}$.

Then (3) gives a suitable criterion of the estimation for regression analysis when the observations have unequal variances are known.

Now, a useful property of natural spline is provided by the following Lemmas.

Lemma 2. (Lyche and Schumaker, 1973) Let $NS^{2m}(t_1,\cdots,t_n)$ denote the collection of all natural splines of order *2m* with knots at the t_j. Let $x_1,\cdots,x_n$ be a basis for $NS^{2m}(t_1,\cdots,t_n)$. Then, there are coefficients $\theta_{0j},\cdots,\theta_{m-1},\delta_{1j},\cdots,\delta_{nj}$ such that

$$x_j(t) = \sum_{i=1}^{m-1}\theta_{ij}t^i + \sum_{i=1}^{n}\delta_{ij}(t-t_i)_+^{2m-1}$$

If $s(t) = \sum_{j=1}^{n}\beta_j x_j(t)$ and $g \in W_2^m[0,1]$, then

$$\int_0^1 g^{(m)}(t)s^{(m)}(t)dt = (-1)^m(2m-1)!\sum_{i=1}^{n}g(t_i)\sum_{j=1}^{n}\beta_j\delta_{ij}$$

Lemma 3. (Eubank, 1988) If $n \geq m$, then the minimizer of (1) is $\hat{\mu}_\lambda = \sum_{j=1}^{n}\beta_{\lambda j}x_j$, where $\beta_\lambda = (\beta_{\lambda 1},\cdots,\beta_{\lambda n})'$ is the solution to

$$(X'X + n\lambda\Omega)\underline{\beta}_\lambda = X'\underline{y} \tag{4}$$

$$\Omega = \{\int_0^1 x_i^{(m)}(t)\cdot x_j^{(m)}(t)dt\}_{i,j=1,n}$$

and $X = \{x_j(t_i)\}_{i,j=1,n}$

Using (4) the vector of fitted values is seen to have the form

$$\underline{\hat{\mu}}_\lambda = (\mu_\lambda(t_1),\cdots,\mu_\lambda(t_n)) = H(\lambda)\underline{y},$$

where $H(\lambda) = X(X'X + n\lambda\Omega)^{-1}X'$.

It is clear that $H(\lambda)$ is symmetric and positive definition matrix. Concerning to the minimization of criterion (3), the following Lemma is needed when errors are heteroscedastic.

Lemma 4. (Eubank, 1988) Let $x_1,\cdots,x_n$ be a basis for $NS^{2m}(t_1,\cdots,t_n)$ and

suppose that $n \geq m$. For fixed $0 < \lambda < \infty$, there is unique minimizer, $\hat{\underline{\mu}}_\lambda$, of (3) in $W_2^m[0,1]$. Moreover, $\hat{\mu}_\lambda \in NS^{2m}(t_1,\cdots,t_n)$ and therefore has the form $\sum_{j=1}^{n}\beta_{\lambda j}x_j$. The coefficients $\underline{\beta}_\lambda = (\beta_{\lambda 1},\cdots,\beta_{\lambda n})'$ are the solution to

$$(X + n\lambda G_w)\underline{\beta}_\lambda = \underline{y} \tag{5}$$

where $G_w = \{(-1)^m(2m-1)!\delta_{ij}/(1+w_j)\}_{i,j=1,n}$.

Multipling X' in (5), we obtain alternative form of (4) by

$$(X'X + n\lambda X'G_w)\underline{\beta}_\lambda = X'\underline{y}.$$

Here $X'G_w$ has typical element

$$\{(-1)^m(2m-1)!\sum_{r=1}^{n}x_r(t_i)\delta_{ij}/(1+w_j)\}_{i,j=1,n}$$

This is just as $W^{-1}\Omega$, where $W = diag(1+w_1,\cdots,1+w_n)$ and $\Omega = \{\int_0^1 x_i^{(m)}(t)\cdot x_j^{(m)}(t)dt\}_{i,j=1,n}$.

Hence we have the following theorem.

Theorem 5. If $n \geq m$, then the minimizing of (3) is $\hat{\mu}_\lambda = \sum_{j=1}^{n}\beta_{\lambda j}x_j$ where $\underline{\beta}_\lambda = (\beta_{\lambda 1},\cdots,\beta_{\lambda n})'$ is the solution to $(X'X + n\lambda X'G_w)\underline{\beta}_\lambda = X'\underline{y}$. And the vector of fitted values is to be the form

$$\hat{\underline{\mu}}_\lambda = (\hat{\mu}_\lambda(t_1),\cdots,\hat{\mu}_\lambda(t_n)) = H(\lambda,\underline{w})\underline{y},$$

where $H(\lambda,\underline{w}) = X(X'X + n\lambda W^{-1}\Omega)^{-1}X'$

$$= X(X'WX + n\lambda\Omega)^{-1}X'W. \tag{6}$$

Let the variance of the jth response be perturbed to $-1 \leq w_j < \infty$, $j = 1,2,\cdots,n$, and let point $\underline{w}_0 = (0,0,\cdots,0)$ represents no perturbation. Then the variance matrix has $W = diag(1+w_j)$ form, and GCV choose $\hat{\lambda}$ to minimize

$$GCV(\lambda,\underline{w}) = \frac{(\underline{y}+\underline{w})^T(I-H(\lambda,\underline{w}))^T(I-H(\lambda,\underline{w}))(\underline{y}+\underline{w})}{\{tr[I-H(\lambda,\underline{w})]\}^2}$$

where $H(\lambda,\underline{w})$ is hat matrix resulting from (6).

Also we can has the local influence measure by the approximate equation when the errors are heteroscedastic,

$$d_{\max}(\sigma_w) \approx -\frac{\partial^2 GCV(\lambda, \underline{w})}{\partial \underline{w}^T \partial \lambda}\Big|_{\hat{\lambda}\, \underline{w}_0}$$
$$\approx tr^2(A)\{(I-2H(\hat{\lambda}))\underline{e}_{\hat{\lambda}} \otimes H(\hat{\lambda})\underline{e}_{\hat{\lambda}} + (\underline{y}+4\underline{e}_{\hat{\lambda}}) \otimes B\underline{e}_{\hat{\lambda}}\}$$
$$+2tr(A)\{tr(B)[(3I-2H(\hat{\lambda})) \otimes H(\hat{\lambda}))\underline{e}_{\hat{\lambda}} - (2I-H(\hat{\lambda}))\underline{y} \otimes \underline{e}_{\hat{\lambda}}]$$
$$-\|\underline{e}_{\hat{\lambda}}\|^2 [\underline{diag(C)}] + 2(\underline{e}_{\hat{\lambda}} H(\hat{\lambda})\underline{e}_{\hat{\lambda}})[\underline{diag(B)}]\}$$
$$-6tr(B)\|\underline{e}_{\hat{\lambda}}\|^2 [\underline{diag(B)}],$$

where $A = I - H(\hat{\lambda})$, $B = H(\hat{\lambda})(I - H(\hat{\lambda}))$,
$C = H(\hat{\lambda})(I - H(\hat{\lambda}))(I - 2H(\hat{\lambda}))$

$[\underline{diag(\cdot)}]$ is the vector that has elements as diagonal values of given matrix, and $\otimes$ is Hadamard product.

4. Example

In order to application of the measure in section 3, we consider the German hyperinflation data. This data consists of values for the logarithm of the money supply as a function of the logarithm for the premium or discount, on a forward contract for foreign exchange during the German hyperinflation. This data has also been studied in the context of spline smoothing by Eubank(1988).

Figure 1. An Index Plot of Residuals for Hyperinflation Data.

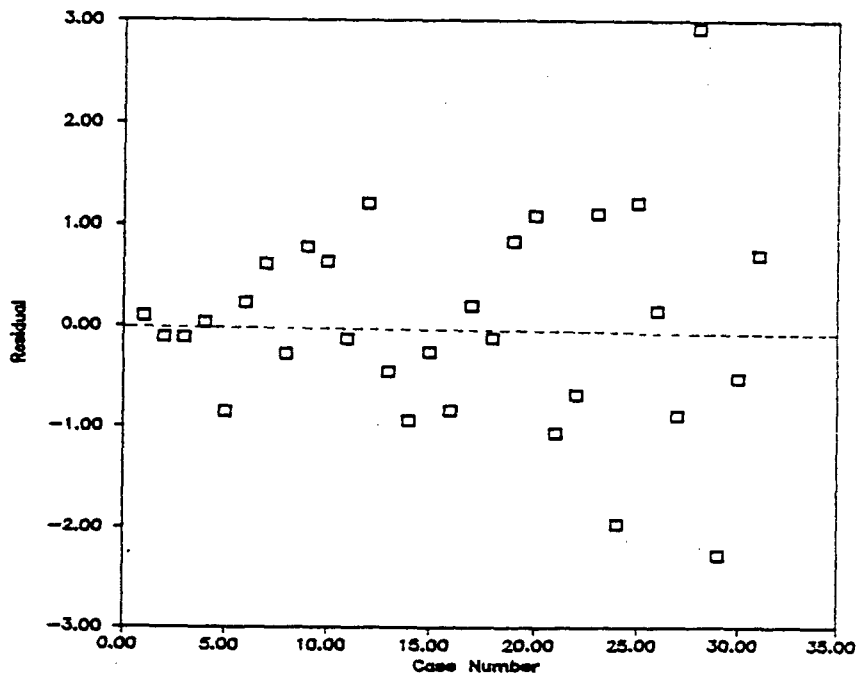
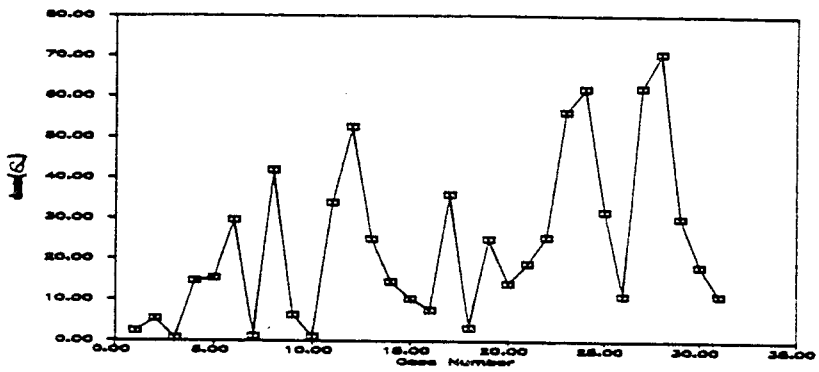


Figure 2. An Index Plot of Diagnostic Measure $d_{\max}(\sigma_w)$.



A plot of the residuals has been presented in Figure 1. Hence we sure that the data set is heteroscedastic. An index plot of the diagnostic values for influence through $d_{\max}(\sigma_w)$, is given in Figure 2. And criteria plot of this measure is given in Figure 3. The cases with the largest absolute components (12, 23, 24, 27, 28) have greatest local influence with respect to heteroscedasticity weights. We obtain the smoothing parameter estimate, $\hat{\lambda}_d = 5.5*10^{-3}$. In this data, we obtain the cases with largest Cook's distance, (19, 24, 28, 29, 31) and smoothing parameter estimate, $\hat{\lambda}_c = 1.1*10^{-7}$. When we suppose homogeneity of errors, $\hat{\lambda}_0$ is $2.42*10^{-4}$. Hence we suspect that Cook's distance made too small estimate and thus is absurd in this case.

Figure 3. Box-Plot as Cut-Off value for $d_{\max}(\sigma_w)$
((2) : two obserbations)

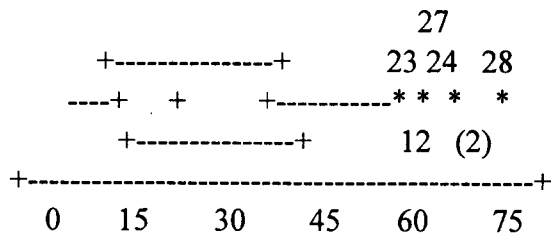
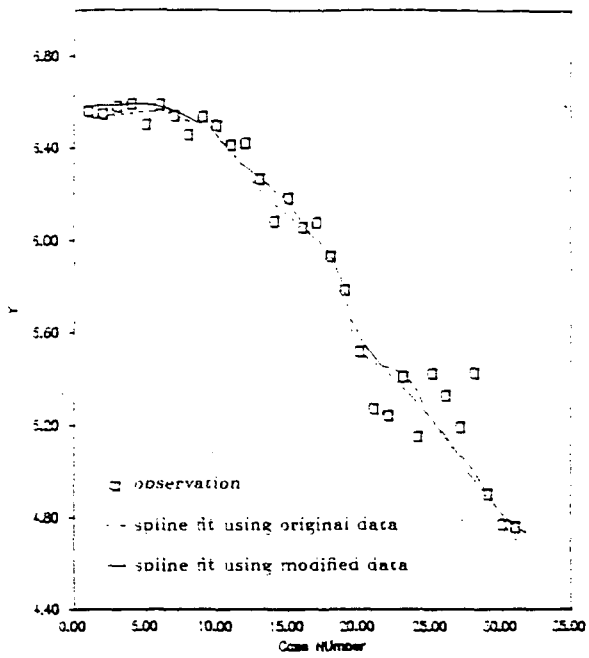


Figure 4 shows the spline fits using $\hat{\lambda}_0$, $\hat{\lambda}_d$ and $\hat{\lambda}_c$ respectively Here, we can observe that smoothing spline fit under the diagnostic procedure is less wiggle.

Figure 4. Spline fits using the original and diagnostic smoothing parameter estimates, respectively.



References

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

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

Eubank, R.L. and Thomas, W.(1993). Dectecting Hetero- scedasticity in Nonparametric Regression, *Journal of the Royal Statistical Society*, Vol. 55, 145-155.

Hall, P. and Tittrington, D.M.(1987). Common Structure of Techniques for Automatic Smoothing Parameters in Regression Problems, *Journal of the Royal Statistical Society*, B, Vol. 49, 184-198.

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

Li, K.C.(1985). From Stein's Unbiased Risk Estimates to the Method of

Generalized Cross-Validation, *Annals of Statististics* Vol. 13, 1352-1377.

Lyche, T. and Schumaker, L.L.(1973). Computation of smoothingand interpolating natural splines via local bases. SIAM J. Numer. Anal. 10, 1027-1038.

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, 1-52.