Journal of the Applied Mathematics, Statistics and Informatics (JAMSI), **3** (2007), No. 1

Confidence interval for the Gasser-Müller estimator

JITKA POMĚNKOVÁ

Abstract

Nonparametric estimates, where kernel smoothing belongs, provide one possible way to find and describe structure in data. The idea of the kernel smoothing can be applied to a simple fixed design regression model and a random design regression model. This article is focused on confidence interval for the kernel estimator with using special type of estimator, the Gasser-Müller one, for fixed design regression model. At the end of this article figures for illustration of described confidence interval on real data are attached.

Mathematics Subject Classification 2000: 62G08 62G15
General Terms: Kernel smoothing, Confidence interval

Additional Key Words and Phrases: Gasser-Müller estimator, nonparametric estimates

1. INTRODUCTION

Nonparametric estimates are one group of method suitable for trend modelling. Kernel smoothing, belonging to this group, produce more flexible estimates than the parametric methods and parametric estimate. They also do not require the knowledge of the distribution data file set. Thus, resultant estimate is free of influence of the distribution type and errors, which can have negative impact in the case of hypothesis testing.

Nonparametric method can be applied on data files with missing values. It is possible, using kernel smoothing, estimate missing values or construct estimates in selected point using whole data file. Also it is possible to construct estimate of the trend derivation, estimate of the break point (increase, decrease). Estimation of confidence interval construction is very often important from analyst point of view.

Convolution type of kernel estimates, Gasser-Müller estimator, gives possibility to easily construct estimate of the function as well as estimate of derivation of the function. It can be applied to a simple fixed design model and a random design regression model.

This article is focused on confidence interval for the kernel estimator with using special type of estimator, the Gasser-Müller one, for fixed design regression model. The aim of this paper is derivation of the formula for construction the Gasser-Müller estimator confidence interval and its usage in case of interpretation. The confidence interval determines area where regression model proceed with given probability and provides for better interpretation of the found regression model. For specification results, given by using the Gasser-Müller estimate and by its confidence interval, the study is completed by an estimate of the first derivation of trend function and its confidence interval. All results are compared. At the end of this article figures for illustration of described confidence interval on real data are attached. Real data set consists of January average temperatures measured in Basel 1755 - 1855.

2. KERNEL SMOOTHING

Let $(x_i, Y_i)_{i=1}^n$, $n \in N$ be a sequence of observations (x, Y), where $x \in R$ and Y is real or simulated measurement. If the values of exogenous variables x are not randomly choosen, we talk about a fixed design regression model. The dependency of value Y on value x can be described by the regression function in the form

$$Y_i = m(x_i) + \epsilon_i, \quad i = 1, \cdots, n, \tag{1}$$

where m is an unknown regression function, $x_i \in [0,1]$ is a point of plan, Y_i is an observation $E(\epsilon_i) = 0 \quad i = 1, \ldots, n$, $D(\epsilon_i) = \sigma^2 > 0 \quad i = 1, \ldots, n$ hold. The values $x_i, i = 1, \ldots, n$ can be described by non-negative integrable function f

$$\int\limits_{-\infty}^{+\infty} f(x)dx = 1; \quad \int\limits_{-\infty}^{x_i} f(x)dx = \frac{i-1}{n-1}; \quad i = 1, \ldots n$$

Definition 2.1. *Let ν,k be non-negative integers, $0 \le \nu < k$. Let $K \in Lip[-1,1]$, support $(K) = [-1,1]$. Let the following moment conditions be satisfied*

$$\int_{-1}^{1} x^j K(x)dx = \begin{cases} 0 & 0 \le j < k, j \ne \nu \\ (-1)^\nu \nu! & j = \nu \\ \beta_k & j = k \end{cases}, \tag{2}$$

Then the function K is called kernel. If $\beta_k \ne 0$ we say that K is kernel of order (ν, k) and write $K \in S_{\nu,k}$.

Definition 2.2. *Let ν,k be non-negative integers, $0 \le \nu < k$, $\mu \ge 1$.Function $K \in C^\mu[-1,1]$, support $(K) = [-1,1]$, which satisfied conditions*

$$\begin{array}{l} (i)\ K^{(j)}(-1) = K^{(j)}(1) = 0 \qquad j = 0, \ldots, \mu - 1 \\ (ii)\ \int_{-1}^{1} x^j K(x)dx = \begin{cases} 0 & 0 \le j < k, j \ne \nu \\ (-1)^\nu \nu! & j = \nu \\ \beta_k \ne 0 & j = k \end{cases}. \end{array} \tag{3}$$

is called kernel of smothness μ, order (ν, k) and write $S^\mu_{\nu,k}$.

The general formula for a kernel estimator can take the following form

$$\hat{m}(x) = \sum_{i=1}^{n} W_i(x,h) Y_i,$$

where $W_i(x,h)$ are weight functions depending on h, i, x and K, $h = h(n)$ is a positive constant and K is a kernel. Denote $K_h(\cdot) = \frac{1}{h} K\left(\frac{\cdot}{h}\right)$.

3. GASSER-MÜLLER ESTIMATOR

Let us have a fixed design regression model described in Section 1. Let us consider the Gasser-Müller estimator

$$\hat{m}^{(\nu)}(x) = \sum_{i=1}^{n} Y_i \frac{1}{h^{\nu+1}} \int_{s_{i-1}}^{s_i} K\left(\frac{x-u}{h}\right) du = \sum_{i=1}^{n} Y_i W_i(x,h). \tag{4}$$

The points of plan x_i, $i = 1,\dots,n$ are ordered according to the size and points s_i, $i = 0,\dots,n$, $s_0 = 0$, $s_n = 1$, $s_i = \frac{x_{i+1}+x_i}{2}$ hold. Thus the weight functions $S_{ij} = W_i(x_j,h)$, $i,j = 1,\dots,n$ in a point of plan x_j, with a bandwith h for an estimator of function $\hat{m}(x)$ take the form

$$W_i(x_j,h) = \frac{1}{h^{\nu+1}} \int_{s_{i-1}}^{s_i} K\left(\frac{x_j-u}{h}\right) du. \tag{5}$$

For the detail see [7], [5].

THEOREM 3.1. *Let us consider the fixed design regression model and let following conditions be satisfied*

1. $m \in C^k([0,1]), k \in N$

2. $K \in \mathcal{M}_{\nu,k}$

3. $\lim_{n\to\infty} h = 0$ *a* $\lim_{n\to\infty} nh^{2\nu+1} = \infty$

Then for every $x \in (0,1)$ *the estimator*

$$\hat{m}^{(\nu)}(x) = \frac{1}{h^{\nu+1}} \sum_{i=1}^{n} Y_i W_i(x,h) = \frac{1}{h^{\nu+1}} \sum_{i=1}^{n} Y_i \int_{s_{i-1}}^{s_i} K\left(\frac{x-u}{h}\right) du \tag{6}$$

is a consistent estimator of $m^{(\nu)}(x)$. *The following formula holds for the variance of this estimation*

$$var(\hat{m}^{(\nu)}(x)) = \frac{\sigma^2}{nh^{2\nu+1}}(C_K + o(1)), \tag{7}$$

where

$$C_K = \int_{-1}^{1} K^2(x)dx, \tag{8}$$

and bias can be expressed

$$E\hat{m}^{(\nu)}(x) - m^{(\nu)}(x) = h^{k-\nu} m^{(k)}(x) B_K + O([nh]^{-1}) + o(h^{k-\nu}), \tag{9}$$

where

$$B_K = (-1)^k \frac{\beta_k}{k!}, \quad \beta_k = \int_{-1}^{1} x^k K(x)dx. \tag{10}$$

PROOF. For the proof see [6]. □

COROLLARY 3.2. *Let Theorem 3.1 hold. Thus*

$$\frac{\hat{m}^{(\nu)}(x) - E\hat{m}^{(\nu)}(x)}{\sqrt{var\ \hat{m}^{(\nu)}(x)}} \longrightarrow N(0,1).$$

PROOF. The proof arise from the Theorem 3.1. For the detail you can see [4]. □

COROLLARY 3.3. *Let Theorem 3.1 hold and let bias be expresed in the form (9). Provided that* $\lim_{n\to\infty} nh^{2k+1} = d^2 > 0$ *for some* $d > 0$,

$$(nh^{2\nu+1})^{1/2} \cdot (\hat{m}^{(\nu)}(x) - m^{(\nu)}(x)) \longrightarrow N(dm^{(k)}(x)B_K, \sigma^2 C_K). \tag{11}$$

PROOF. The proof is straightforward consequence of the Theorem 3.1. and Corollary 3.2.. □

Denote u_α a number (a quantil) oversteped with probability α by standard normal distributed random variable, ie. $1 - \Phi(u_\alpha) = \alpha$. We are searching for interval which overstep an estimate $m^{(\nu)}(x)$ with probability close to $1-\alpha$. This interval has been called confidence interval.

THEOREM 3.4. *Let Theorem 3.1, Corollary 3.2. and Corollary 3.3. hold. Thus confidence interval for the Gasser-Müller estimate of the function* $m^{(\nu)}(x;h)$ *takes the form*

$$\hat{m}^{(\nu)}(x;h) \pm \xi, \quad \xi = u_{1-\frac{\alpha}{2}} \sqrt{\frac{C_K \sigma^2(x)}{nh^{2\nu+1}}}. \tag{12}$$

PROOF. Let Theorem 3.1, Corollary 3.2. and Corollary 3.3. hold. The confidence interval is deduced from relation

$$P\left(-u_{\alpha/2} \leq \frac{\sqrt{nh^{2\nu+1}}(\hat{m}^{(\nu)}(x) - m^{(\nu)}(x)) - dm^{(k)}(x)B_K}{\sqrt{\sigma^2 C_K}} \leq u_{1-\alpha/2}\right) \approx 1 - \alpha.$$

After the derivation is obtained

$$P\left(\hat{m}^{(\nu)}(x) - \frac{dm^{(k)}(x)B_K}{\sqrt{nh^{2\nu+1}}} - u_{1-\alpha/2}\sqrt{\frac{\sigma^2 C_K}{nh^{2\nu+1}}} \leq m^{(\nu)}(x) \leq \right.$$

$$\left. \leq \hat{m}^{(\nu)}(x) - \frac{dm^{(k)}(x)B_K}{\sqrt{nh^{2\nu+1}}} + u_{1-\alpha/2}\sqrt{\frac{\sigma^2 C_K}{nh^{2\nu+1}}}\right) \approx 1 - \alpha. \tag{13}$$

We suppose negligible bias, we might neglect the second member in formula (12) as well. □

For obtaining the variance at the fixed points $x \in [0,1], i = 1, \ldots, n$ we can use the formula

$$var(\hat{m}(x)) \approx \frac{C_K \sigma^2(x)}{nh^{2\nu+1}}, \tag{14}$$

where for estimation $\sigma^2(x)$ we use the formula

$$\hat{\sigma}^2(x) = \sum_{i=1}^{n} W_i(x,h)(Y_i - \hat{m}(x))^2. \tag{15}$$

In both cases we can use formula (15) for an estimation bias $\sigma^2(x)$ (see [1]).

Algorithm

Algorithm for the confidence interval construction of the Gasser-Müller estimate can by done as follows:

step 1: calculate Gasser-Müller estimator $\hat{m}^{(\nu)}(x)$ using formula (4)

step 2: estimate value $\sigma^2(x)$ using formula (15)

step 3: calculate ξ for given estimation and $100(1-\alpha)\%$ confidence interval using formula (12)

4. APPLICATION

The algorithm described in the preceding part was used for the construction of confidence interval for real data set - January average temperature measured in Basel during the period 1755-1855. The lowest temperature was measured in 1830 and it was $-8.8°C$, the highest in 1834 and it was $5.4°C$. Description of label x is equidistantly distributed on interval $[0,1]$ according time period 1755-1855 for easier calculation.

For an estimation of the Januare average temperature the kernel

$$K(x) = -\frac{105}{256}(x^2-1)(33x^4-30x^2+5), \quad K \in \mathcal{S}^1_{0,6}$$

$\alpha = 0.05$ and the bandwidth $h = 0.65$ were used.

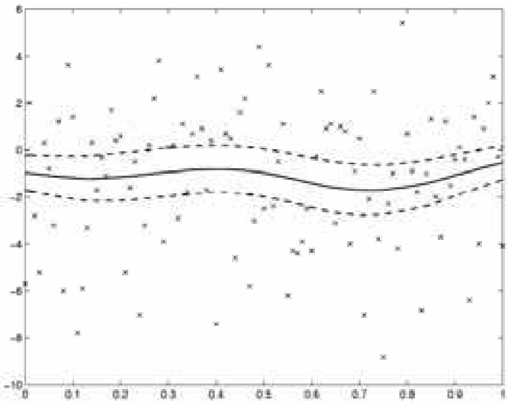

Fig. 1. The confidence interval for Basel 1755 - 1855 January temperature (Dashed line is the confidence interval, solid line is estimation of January average temperatures).

For an estimation of the first derivative of January average temperature the kernel

$$K(x) = \frac{15}{16}(1-x^2)^2, \quad K \in \mathcal{S}^2_{0,2}$$

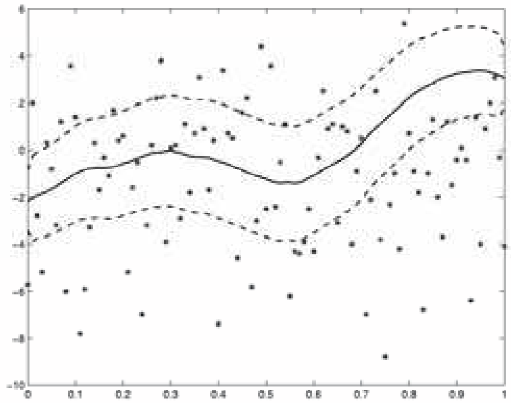

Fig. 2. The confidence interval for an estimation of the first derivative function describing January average temperatures from Basel 1755 - 1855 (Dashed line is the confidence interval, solid line is estimation of the first derivative of January average temperatures).

$\alpha = 0.05$ and $h = 0.45$ were used.

In both cases smooth kernels were used. In case of estimate January average temperature optimum kernel (smoothness $\mu = 1$) was used. This type of kernel minimize average mean square error. In second case, first derivation of January average temperature estimate was chosen kernel with higher smoothness ($\mu = 2$). Choosing optimum bandwidth h was done using cross-validation method. For the detail of choosing kernel type and bandwidth have a look [1], [6].

Presented figures show real data set with Gasser-Müller estimate of the trend in given data structure (Fig. 1) and the first derivation trend of same data structure (Fig. 2). Result of the first derivation trend estimate in will be used for specification of estimated trend.

From the figure of the first derivation is evident one stationary point (x=0,7; year 1825). In the interval behind this point values of the first derivation are positive, so we can expect that function is positive in the period after stationary point. If we have a look on the fig. 2, point (x=0.3; year 1785) may appear as stationary point as well. Value of the first derivation in this point is not unfortunately equal to zero, but it can leads to an idea that from estimate construction and specially corresponding confidence interval may exist stationary point in the area of the point (x=0.3; year 1785). However, character of the first derivation function progress indicate that function on interval 1755 – 1825 (into stationary point discussed above; x = 0.3; year 1825) will be rather decreasing.

Knowledge from the Fig. 2 employ for specification estimated trend of the January average temperatures measured in Basel 1755 - 1855 in Fig. 1 and for its interpretation. Between years 1755 - 1825 (Fig. 1) January average temperatures in Basel have variable trend with slight decrease followed by slight increase and after that again slight decrease. Data set character shows volatility. From estimate of the trend may arise an idea, that in given period (1755 - 1825) are two extreme points - minimum and maximum.

If we compare this indication of the progress with figure of the first derivation, we can see, that between years 1755 - 1825, the trend is really decreasing. Founded confidence interval (respectively area where regression model proceed with given probability) of the January average temperatures trend supposes this idea. For its support analysis of the results from the first derivation trend (discussed above) is used.Farther in the period 1825 - 1855 increasing trend of the January average temperatures is expected. This fact is validates by the first derivation trend chart.

From the long-term point of view we can say, that the January average temperatures measured in Basel 1755 - 1855 was decreasing in the period 1755 - 1825 and were increasing in the period 1825 - 1855. This fact can be very useful in agribusiness, when temperature analysis can help to increase effectiveness of agriculture work.

5. ACKNOWLEDGMENTS

This work is part of research project "The Czech Economy in the Process of Integration and Globalisation, and the Development of Agricultural Sector and the Sector of Service under the New Conditions of the Integrated European Market".

REFERENCES

(1) Härdle W.: Applied nonparametric regression. Cambridge University Press, Cambridge, 1990,pp. 71-75.

(2) Härdle, W. , Schimek, M. G.: Statistical Theory and Computational Aspects of Smoothing, 1994.

(3) Horová, I.: Some Remarks on Kernels, Journal of Computational Analysis, Vol.2., No.2.,2000, pp. 253-263. Zbl 0964.62031

(4) Müller, H.-G.: Nonparametric Regression analysis of Longitudinal data. Lecture Notes in Statistics 46. Springer - Verlag, 1988, pp. 26-111.

(5) Poměnková, J. Gasser-Müller estimate. Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis. 2004. sv. L II, No 3, pp. 167-176. ISSN 1211-8516.

(6) Poměnková J.: Některé aspekty vyhlazováni regresní funkce, PhD-thesis, Ostrava, 2005.

(7) Wand, M. P., Jones, M. S.: Kernel Smoothing. Chapman & Hall, London,1995.

Jitka Poměnková,
Mendel University of Agriculture and Forestry Brno,
Faculty of Business and Economics,
Zemědělská 1, 621 00 Brno, Czech Republic
e-mail: pomenka@mendelu.cz

Received July 2007