Bulletin of National University of Uzbekistan: Mathematics and Natural Sciences

Volume 1 | Issue 4 Article 3

9-9-2018

The optimal risk of estimator of conditional distribution function in a model of heteroscedastic regression with weakly dependent observations

Farkhad Abdikalikov
Karakalpak State University, abdikalikovf@mail.ru

Laura Reymova
Karakalpak State University, reymova1998@gmail.com

Volume 1, Issue 4 (2018)

THE OPTIMAL RISK OF ESTIMATOR OF CONDITIONAL DISTRIBUTION FUNCTION IN A MODEL OF HETEROSCEDASTIC REGRESSION WITH WEAKLY DEPENDENT OBSERVATIONS

Abdikalikov F. A., Reymova L.
Karakalpak State University, Nukus, Uzbekistan
e-mail: abdikalikovf@mail.ru, reymova1998@gmail.com

Abstract

Paper is devoted to estimation of conditional distribution function in heteroscedastical regression model, in which responses are α-mixing random variables. It is found the expression for mean square deviation of estimator and optimal window width sequence.

Keywords: model of heteroscedastical regression, α-mixing, kernel estimate.

Mathematics Subject Classification (2010): 46N30, 62H12.

Introduction

The model of heteroscedastical regression of the variable of response Z on X is determined by the formula [1]:

$$Z = m(X) + \sigma(X)\varepsilon,$$

where X - is a random covariate, random error ε - is independent of X. The functionals $m(x) = M\left(Z/X = x\right)$ and $\sigma^2(x) = D\left(Z/X = x\right)$ respectively, are conditionally mean and variance functions of regression Z on X that is, possible heterossedality. Define the conditional distribution function (d.f.) $F_x(t) = P(Z \leq t/X = x)$, $(t; x) \in R^+ \times D_X$. By definition, functionals $m(x) = T(F_x(\cdot))$ and $\sigma(x) = S(F_x(\cdot))$ for all $a \geq 0$ and $b \in R$ satisfy the equalities (see, [1]):

$$T\left(Q_{aZ+b}(\cdot/x)\right) = aT\left(Q_Z(\cdot/x)\right) + b = am(x) + b,$$

$$S\left(Q_{aZ+b}(\cdot/x)\right) = aS\left(Q_Z(\cdot/x)\right) = a\sigma(x),$$

where $Q_{aZ+b}(t/x) = P(aZ + b \leq t/X = x) = F_x\left(\frac{t-b}{a}\right)$. For them, it is true representations

$$m(x) = \int_0^1 F_x^{-1}(s)J(s)ds, \;\; \sigma^2(x) = \int_0^1 \left(F_x^{-1}(s)\right)^2 J(s)ds - m^2(x), \qquad (1)$$

where $F_x^{-1}(s) = \inf\{y : F_x(y) \geq s\}$, $0 \leq s \leq 1$, quantile function of a random variable (r.v.) Z for a given $X = x$ and $J(s)$ - is a given score function such that $J(s) \geq 0$ and $\int_0^1 J(s)ds = 1$. As $J(s)$ one can take a function $J(s) = I(0 \leq s \leq 1)$.

1 Preliminaries

The basic problem is to estimate the conditional d.f. $F_x(t)$ for a sample of observations on a pair (Z, X): (Z_1, X_1), $(Z_2, X_2), ..., (Z_n, X_n)$ where the subsample $\{Z_1, \dots, Z_n\}$ satisfies the following condition of α-mixing at $n \to \infty$:

$$\alpha(n) = \left\{ \sup_{k \geq 1} |P(A \cap B) - P(A)P(B)| : A \in \mathcal{F}_1^k(Z),\, B \in \mathcal{F}_{k+n}^{\infty}(Z) \right\} \to 0,$$

where $\mathcal{F}_i^k(Z)$ - σ - is the algebra of events generated by a set of a r.v. $\{Z_j,\, i \leq j \leq k\}$. Among most classes of dependent r.v. α-mixed r.v. are most common in practice (see, [2]). It should be noted that indicators $\{I(Z_i \leq t)\}$ also form a sequence α-mixingled r.v.-s.

As an estimator, for $F_x(t)$, we consider the following Nadarya-Watson statistics (see, [3, 4]):

$$F_{xh}(t) = \sum_{i=1}^{n} \Psi_{ni}(x; h_n)\, I(Z_i \leq t), \quad (t; x) \in R^+ \times D_X\,, \tag{2}$$

where weights

$$\Psi_{ni}(x; h_n) = \left(\sum_{j=1}^{n} k\left(\frac{x - X_j}{h_n} \right) \right)^{-1} k\left(\frac{x - X_i}{h_n} \right), \quad i = 1, \dots, n,$$

are given by the sequence $h_n \downarrow 0$ as $n \to \infty$ and by the kernel $k(\cdot)$. It is easy to see by direct calculation that the conditional expectation and variance of the estimate (2) for given $X_1, ..., X_n$ are equal

$$\begin{aligned} M^* F_{xh}(t) &= M\left[F_{xh}(t) / X_1, ..., X_n\right] = \sum_{i=1}^{n} \Psi_{ni}(x; h_n)\, F_{X_i}(t), \\ D^* F_{xh}(t) &= D\left[F_{xh}(t) / X_1, ..., X_n\right] = \sum_{i=1}^{n} \Psi_{ni}^2(x; h_n)\, F_{X_i}(t)\left(1 - F_{X_i}(t)\right). \end{aligned} \tag{3}$$

Since for sufficiently large n and $x \in (h_n, 1 - h_n)$, $F_{X_i}(t) \approx F_x(t)$ and $\sum_{i=1}^{n} \Psi_{ni}(x; h_n) = 1$, it is natural to expect that the right sides of formulas (3) are asymptotically unbiased estimates of the corresponding expectations, that is, for $n \to \infty$ and $(t, x) \in R^+ \times D_X$:

$$MF_{xh}(t) = F_x(t) + o(1), \quad DF_{xh}(t) = \frac{1}{nh_n} F_x(t)\left(1 - F_x(t)\right) + o(1). \tag{4}$$

In this paper, we will estimate the quadratic risk of the estimate (2). We denote $\dot{F}_x(t) = \frac{\partial F_x(t)}{\partial x}$, $\ddot{F}_x(t) = \frac{\partial^2 F_x(t)}{\partial x^2}$. For a sequence $\{h_n, n \geq 1\}$, a kernel $k(\cdot)$ and a conditional distribution function F_x, we introduce the conditions (see, [3]):

(C1) At $n \to \infty$, $h_n \downarrow 0$ and $nh_n \to \infty$;

Volume 1, Issue 4 (2018)

(C2) The kernel $k(\cdot)$ is a continuous, bounded and symmetric density with a compact support $[-M,\, M]$, for$M>0$;

(C3) For a fixed $t\in R^{+}$ derivative $\ddot{F}_x(t)$ exists and is bounded in a neighborhood U_x of x.

Further, we also need the following statement.

Suppose ξ and η are two r.v. that measurable with respect to $\mathcal{F}_1^k(Z)$ and $\mathcal{F}_{k+n}^{\infty}(Z)$, respectively, and

$$\|\xi\|_\infty = ess-\sup|\xi| = \inf\{t\in R^{+}\cup\{+\infty\} : P(|\xi|>t)=0\}.$$

Lemma 1 ([2]). *It is true the following inequality*

$$|Cov(\xi,\eta)| \le 4\alpha(n)\|\xi\|_\infty\cdot\|\eta\|_\infty.$$

For example, for an indicator $I(Z_i\le t)$:

$$ess-\sup|I(Z_i\le t)|=1.$$

2 Main part

The following statement gives an estimate for the standard deviation of $F_{xh}(t)$ from $F_x(t)$.

Theorem 1. *Let the conditions (C1)-(C3) and the sequence $\{Z_i, i\ge 1\}$ are satisfied the condition of α-mixing so that, for $n\to\infty$, $\alpha(n)\to 0$ and*

$$\frac{1}{nh_n^2}\sum_{i=1}^{n-1}\alpha(i)\to 0.$$

Then for $n\to\infty$

$$\begin{aligned} M[F_{xh}(t)-F_x(t)]^2 &= \left[\tfrac{h_n^2}{2}\ddot{F}_x(t)\int u^2k(u)du\right]^2+ \\ &+ \tfrac{1}{nh_n}\int k^2(u)du\cdot F_x(t)(1-F_x(t))+o\left(h_n^4+\tfrac{1}{nh_n}\right). \end{aligned}$$

Proof. Since

$$M[F_{xh}(t)-F_x(t)]^2 = DF_{xh}(t)+[MF_{xh}(t)-F_x(t)]^2, \tag{5}$$

we estimate the terms on the right side of the last equality separately. Denote $\xi_i=\frac{x-X_i}{h_n}$. Then, by the Taylor expansion, we easily have the equalities

$$F_{X_i}(t)=F_{x-h_n\xi_i}(t)=F_x(t)-h_n\xi_i\dot{F}_x(t)+\frac{h_n^2}{2}\xi_i^2\ddot{F}_x(t)+o(h_n^2), \tag{6}$$

$$\begin{aligned} F_{X_i}^2(t)=F_{x-h_n\xi_i}^2(t) &= F_x^2(t)-2h_n\xi_iF_x(t)\dot{F}_x(t)+ \\ &+ \tfrac{h_n^2}{2}\xi_i^2\left(\dot{F}_x(t)\right)^2+\tfrac{h_n^2}{2}\xi_i^2F_x(t)\ddot{F}_x(t)+o(h_n^2). \end{aligned} \tag{7}$$

Taking into account (6) and the symmetry of the kernel $k(\cdot)$ (in condition (C2)) for the expectation, we have

$$\begin{aligned} MF_{xh}(t) &= F_x(t) + \frac{h_n^2}{2}\ddot{F}_x(t)\frac{\sum\limits_{i=1}^{n}\xi_i^2 k(\xi_i)}{\sum\limits_{i=1}^{n} k(\xi_i)} + o(h_n^2) = \\ &= F_x(t) + \frac{h_n^2}{2}\ddot{F}_x(t)\int u^2 k(u)du + o(1), \end{aligned} \tag{8}$$

where we used following equalities

$$\begin{aligned} R_n^{(0)} &= \frac{1}{nh_n}\sum_{i=1}^{n} k(\xi_i) = \int k(u)du + o(1) = 1 + o(1), \ \ n \to \infty, \\ R_n^{(1)} &= \frac{1}{nh_n}\sum_{i=1}^{n} \xi_i k(\xi_i) = \int uk(u)du + o(1) = o(1), \ \ n \to \infty, \\ R_n^{(2)} &= \frac{1}{nh_n}\sum_{i=1}^{n} \xi_i^2 k(\xi_i) = \int u^2 k(u)du + o(1), \ \ n \to \infty. \end{aligned} \tag{9}$$

From (9), in particular, when $n \to \infty$ follows the asymptotic unbiasedness of estimate $MF_{xh}(t) = F_x(t) + O(h_n^2)$. Now we calculate the conditional variance of the estimate:

$$\begin{aligned} DF_{xh}(t) &= \frac{\sum\limits_{i=1}^{n} k^2(\xi_i)\left[F_{X_i}(t) - F_{X_i}^2(t)\right]}{\left[\sum\limits_{i=1}^{n} k(\xi_i)\right]^2} + \\ &+ \sum_{i=1}^{n}\sum_{\substack{j=1 \\ i\neq j}}^{n} \frac{k(\xi_i)k(\xi_j)Cov\left(I(Z_i \le t), I(Z_j \le t)\right)}{\left[\sum\limits_{i=1}^{n} k(\xi_i)\right]^2} = D_{1n} + D_{2n}. \end{aligned} \tag{10}$$

We study the terms on the right-hand side of (10). Hence, we calculate

$$Q_n^{(0)} = \frac{1}{nh_n}\sum_{i=1}^{n} k^2(\xi_i) = \int k^2(u)du + o(1), \ \ n \to \infty, \tag{11}$$

and also due to the symmetry of $k(\cdot)$,

$$Q_n^{(1)} = \frac{1}{nh_n}\sum_{i=1}^{n} \xi_i k^2(\xi_i) = \int uk^2(u)du + o(1) = o(1), \ \ n \to \infty. \tag{12}$$

Using relations (6), (7), (9), (11) and (12), we have

$$\begin{aligned} D_{1n} &= \left[\sum_{i=1}^{n} k(\xi_i)\right]^{-2}\left\{\left[F_x(t) - F_x^2(t)\right]\sum_{i=1}^{n} k^2(\xi_i) + \right. \\ &+ \left. \frac{h_n^2}{2}\left[\ddot{F}_x(t) - \left(\dot{F}_x(t)\right)^2 - F_x(t)\ddot{F}_x(t)\right]\sum_{i=1}^{n}\xi_i^2 k^2(\xi_i) + o(1)\right\} = \\ &= \frac{1}{nh_n}\left[F_x(t) - F_x^2(t)\right]\int k^2(u)du + o(1). \end{aligned} \tag{13}$$

Volume 1, Issue 4 (2018)

Now we evaluate D_{2n}. Let be $c(j-i) = Cov\left(I(Z_i \le t), I(Z_j \le t)\right)$. Then, in view of the first equality in (9),

$$\begin{aligned} |D_{2n}| &\le \left[R_n^{(0)}\right]^{-2} \frac{1}{n^2h_n^2} \left| \sum_{\substack{i=1 \\ i\neq j}}^{n} \sum_{j=1}^{n} k(\xi_i)k(\xi_j)c(j-i) \right| \le \\ &\le \frac{2}{n^2h_n^2} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} k(\xi_i)k(\xi_j)\left|c(j-i)\right| \le \frac{2k_0^2}{n^2h_n^2} \sum_{i=1}^{n} (n-i)|c(i)| \le \\ &\le \frac{2k_0^2}{nh_n^2} \sum_{i=1}^{n-1} |c(i)| \le \frac{8k_0^2}{nh_n^2} \sum_{i=1}^{n-1} \alpha(i) = o(1),\ n \to \infty, \end{aligned} \tag{14}$$

where $k_0 = \sup\limits_{u\in R} k(u) \in (0,\infty)$, the lemma and conditions of the theorem are used. Now the statement of the theorem follows from (5), (8), (10), and (14). Theorem 1 is proved. □

Corollary 1. *Under the conditions of Theorem 1, for sufficiently large n, we have the following asymptotic representation for the variance of the estimate $F_{xh}(t)$:*

$$DF_{xh}(t) = \frac{1}{nh_n} F_x(t)(1-F_x(t)) \int k^2(u)du + o(1),\ x \in (h_n, 1-h_n). \tag{15}$$

From Theorem 1 also follows for given covariates $X_1, \ldots, X_n$ the mean square consistency of the estimate $F_{xh}(t)$. Under the conditions of Theorem 1, we can write the following asymptotic representation for the quadratic risk of an estimate $F_{xh}(t)$ with a given weight function $w(\cdot)$ for a fixed $t \in R^+$:

$$\int\limits_a^b M[F_{xh}(t) - F_x(t)]^2 w(x)dx \approx I(h_n),$$

where

$$I(h_n) = \frac{RA}{4}h_n^4 + \frac{QB}{nh_n},\ Q = \int\limits_{-M}^{M} k^2(u)du,\ R = \int\limits_{-M}^{M} u^2k(u)du,\ A = \int\limits_a^b \ddot{F}_x(t)w(x)dx,$$

$$B = \int\limits_a^b F_x(t)(1-F_x(t))w(x)dx.$$

In order to find the optimal sequence $\{h_n, n \ge 1\}$ that gives the least value to risk, we solve the equation:

$$\frac{\partial I(h_n)}{\partial h} = RAh_n^3 - \frac{QB}{nh_n^2} = 0,$$

from where we find $h_{n,opt} = Cn^{-1/5}$, where $C = \left(\frac{QB}{RA}\right)^{1/5}$.

For example, if $k(\cdot)$ is the uniform distribution density on $[-1,1]$, then $Q = 1/2$ and $R = 1/3$ and the value C also depends only on the degree of smoothness of the function $F_x(t)$ with respect to $x \in D_X$.

References

[1] Akritas M.G., Van Keilegom I. Nonparametric estimation of the residual estimation. Scandinavian Journal of Statistics, Vol. 28, Issue 3, 549–567 (2001).

[2] Doukhan P. Mixing: Properties and Examples. Lecture Notes in Statistics, Vol. 85, Springer-Verlag, New York, (1994). – 144 pp.

[3] Nadaraya E.A. Nonparametric estimation of probability density and regression curve. Tbilisi Univ. Publishing House, Tbilisi, (1983). – 194 pp. [in Russian].

[4] Watson J.S. Smooth regression analysis. Sankhyā: The Indian Journal of Statistics, Series A, Vol. 26, No. 4, 359–372 (1964).