한국수학교육학회지 <수학교육>
1991. 6. 제 30권 제 1호 47-50

Journal of the Korea Society of Mathematical Education
June 1991: Vol. 30, No.1, 47-50

Nonparametric Estimators for Percentile Regression Functions

Eun-Sook Jee

Kwangwoon University, Seoul, 139-701, Korea

ABSTRACT. We consider the regression model $H = h(x) + E$, where h is an unknown smooth regression function ard E is the random error with unknown distribution F. In this context we present and examine the asymptotic behavior of some nonparametric estimators for the percentile regression functions $\xi_p(x) = h(x) + \xi_p$, where $0 < p < 1$ and $\xi_p = \inf\{x : F(x) \geq p\}$

1. Introduction

Consider the regression model

$$Y_{ij} = h(x_i) + E_{ij} \qquad (j = 1, \ldots, m_i \text{ and } i = 1, \ldots, n),$$

where $h(x)$ is an unknown function defined on the colsed interval $[0,1]$ (or any closed interval), and E_{ij} are independent and identically distributed random varialbes from a population with unknown distribution function $F(x)$. Under this type of regression models, investigators are frequently interested in estimates of different percentilies $\xi_p(x) = h(x) + \xi_p$ of the distribution of Y for a given design point x, where $0 < p < 1$, and $\xi_p = \inf\{x : F(x) \geq p\}$ is the (100p)th percentile of the distribution function F. For example, assuming that the function $h(x)$ is a linear function $\alpha + \beta x$ and E is normally distributed with mean 0 and variance σ^2, the problem of obtaining point estimators and confidence bands for $\xi_p(x) = \alpha + \beta x + \xi_p$ was considered by Easterling (1969), Turner and Bowden (1977), Griffiths and Willcox (1978), among others. Here $\xi_p = \sigma Z_p$, and Z_p is the (100p)th percnetile of the standard normal distribution. In this paper, we are concerned with the problem of estimating the general percentile regression function $\xi_p(x)$ based on the random sample $\{(x_i, Y_{ij}), j = 1, \ldots, m_i, i = 1, \ldots, n\}$ when the functions h and F are both unknown.

Since design points are selected from $[0,1]$, we may without loss of generality assume that $0 = x_0 \leq x_1 \leq x_2 < \cdots < x_n \leq 1$. To define nonparametric estimators of $\xi_p(x)$, $0 < p < 1$, our first step is to estimate the distribution function $F_x(y) = F(y - h(x))$ of Y for a given vaule of x by using

$$\begin{aligned} F_{x,n}(y) = \widehat{G_1}(y) \int_{-\infty}^{x} a_n^{-1} K\left[\frac{x-z}{a_n}\right] dz + \sum_{i=2}^{n-1} \widehat{G_i}(y) \int_{x_{i-1}}^{x_i} a_n^{-1} K\left[\frac{x-z}{a_n}\right] dz \\ + \widehat{G_n}(y) \int_{x_{n-1}}^{\infty} a_n^{-1} K\left[\frac{x-z}{a_n}\right] dz \end{aligned} \tag{1.1}$$

see Stone (1977). Here $\widehat{G_i}(y) = m_i^{-1} \sum_{j=1}^{m_i} I(Y_{ij} \leq y)$ $(i = 1, \ldots, n)$, $I(\cdot)$ is the indicator function, the weight function $K(x)$ is a probability density function vanishing outside some closed interval $[-L, L]$ and bandwidth parameter a_n is a constant tending to 0 as $n \to \infty$. In view of (1.1) we note that $F_{x,n}(y)$ is a right continuous function and increases by jumps only at points Y_{ij}. In addition, if we let $x_n \to 1$ as $n \to \infty$, then for each $x \in (0,1)$, $F_{x,n}(y)$ can be expressed as

$$F_{x,n}(y) = \sum_{i=1}^{n} \{\widehat{G}_i(y) \int_{x_{i-1}}^{x_i} a_n^{-1} K \left[\frac{x-z}{a_n}\right] dz\} \tag{1.2}$$

for all sufficiently large n. In the sequel, we shall always assume that $x_n \to 1$, $n \to \infty$, and hence we can utilize the simple expression (1.2) for $F_{x,n}(y)$.

As we will see in Section 2, the random function $F_{x,n}(y)$ is a good estimator of $F_x(y)$. Thus to estimate $\xi_p(x)$, $0 < p < 1$, we simply consider the intuitive estimator

$$\xi_{p,n}(x) = \inf\{y : F_{x,n}(y) \geq p\}$$

In this paper, some stochastic properties of $\xi_{p,n}(x)$ will be investigated. Specifically, we show that $\xi_{p,n}(x)$ is a consistent estimator of the unknown percentile regression function $\xi_p(x)$. Moreover, $\xi_{p,n}(x)$ is shown to be asymptotically normal under very mild conditions.

In regression analysis, the estimates $\xi_{p,n}(x)$ may furnish very good descriptive statistics. Besides, this estimation procedure has application to discrimination on percentiles in regression; see Easterling (1969) and Steinhorst and Bowden (1971). Moreover, the median regression function estimate $\xi_{1/2,n}(x)$ provides a good estimate for the regression function h, when the distribution function F has median $\xi_{1/2} = 0$. Other competitive estimators were considered by Benedetti (1977), and Cheng and Lin (1981 a, b).

2. Strong Consistency of $\xi_{p,n}(x)$

In the following, we provide a set of sufficient conditions showing that $\xi_{p,n}(x)$ is indeed a consistent estimator of $\xi_p(x)$. Define $\delta_n = \max_{1 \leq i \leq n}(x_i - x_{i-1})$. $N_n = \min_{1 \leq i \leq n} m_i (\geq 1)$ and $||k||_\infty = \sup|K(x)|$. Throughout this paper, we also let c denote a generic constant which may not be the same at each appearance.

Theorem 2.1. *Let $0 < p < 1$ and $x \in (0,1)$. Assume that $F, h \in$ Lip(1), $||k||_\infty < \infty$, $\delta_n \to \infty$, $n \to \infty$, and $\beta_n N_n^{-1} \delta_n a_n^{-1} \log^2 n = 0(1)$, $n \to \infty$, where $\beta_n \to \infty$, is any sequence of positive constants. If ξ_p is the unique solution y of $F(y^{-1}) \leq P \leq F(y)$ then with probability one,*

$$\xi_{p,n}(x) \to \xi_p(x), \qquad n \to \infty$$

Proof: For each $\alpha > 0$ and $x \in (0,1)$,

$$F_x(\xi_p(x) - \alpha) = F(\xi_p - \alpha) < p < F(\xi_p + \alpha) = F_x(\xi_p(x) + \alpha), \tag{2.1}$$

by the uniqueness condition of the theorem. Now using (1.2) and a moment inequality of the exponential form (see, e.g., Lamperti (1966), pp.43–44), we obtain

$$P\{|F_{x,n}(y) - EF_{x,n}(y)| > \alpha\} \leq cn^{-\alpha\beta_n^{1/2}}, \quad \text{for each } \alpha > 0$$

Further, $|EF_{x,n}(y) - F_x(y)| \to 0$, $n \to \infty$. Thus, in view of (2. 1), we may conclude that

$$P\{F_{x,m}(\xi_p(x) - \alpha) < p < F_{x,m}(\xi_p(x) + \alpha), \quad \text{all } m \geq n\} \to 1, \quad n \to \infty.$$

Consequently, for each $\alpha > 0$,

$$P\{\xi_p(x) - \alpha < \xi_{p,m}(x) < \xi_p(x) + \alpha, \text{ all } m \geq n\} \to 1, \quad n \to \infty$$

This finishes the proof.

Remarks:

(i) If we apply the same approach used in Cheng and Lin (1981a, Theorem 3), then for each constant c,

$$\mathrm{Sup}_{X \in [a,b]} |F_{x,n}(\xi_p(x) + c) - F_x(\xi_p(x) + c)| \overset{\text{W.P.1}}{\longrightarrow} 0, n \to \infty$$

where $0 < a \leq b < 1$. According to the above proof, this result will then imply that

$$\mathrm{Sup}_{X \in [a,b]} |\xi_{p,n}(x) - \xi_p(x)| \overset{\text{W.P.1}}{\longrightarrow} 0, n \to \infty$$

(ii) We have the following observations (with regularity conditions omitted):

(1) $\xi_{p,n}(x) - \xi_p(x) = \frac{F_x(\xi_{p,n}(x)) - p}{f_x F_x^{-1}(\theta_{p,n}(x))}$, where $F_x(\xi_{p,n}(x)) \wedge p < \theta_{p,n}(x) < F_x(\xi_{p,n}(x)) \vee p$,

(2) $F_x(\xi_{p,n}(x)) = \inf\{y : F_{x,n}(F_x^{-1}(y)) \geq p\}$

Morevoer, using a theorem by Singh (1975), we have

$$\gamma_n \cdot \mathrm{Sup}_{y \in R} |F_{x,n}(y) - F_x(y)| \overset{\text{W.P.1}}{\longrightarrow} 0, n \to \infty \tag{2.2}$$

with $\gamma_n \to \infty$. $n \to \infty$. Thus under some appropriate conditions in conjunction with (1), (2) and results in Vervaat (1972), (2.2) forces

$$\gamma_n \cdot \mathrm{Sup}_{0 \leq p \leq 1} |\xi_{p,n}(x) - \xi_p(x)| \overset{\text{W.P.1}}{\longrightarrow} 0, n \to \infty$$

(iii) In Theorem 2.1, β_n is any sequence of positive constants tending to ∞ as $n \to \infty$. Thus if $m_i = 1$, $i = 1, \ldots, n$, and $\delta_n = n^{-1}$, we may let $\beta_n = \log\log n$ and then choose $a_n = cn^{-1}\log^2 n \log\log n$, where c is any fixed positive constant.

References

1. Benedetti, J. O., *On the nonparametric estimation of regression function*, J. Roy. Statist. Soc. B **39** (1977), 248–253.
2. Cheng, K. F. and Lin, P. E., *Nonparametric estimation of a regression function*, Z. Wahrscheinlichkeitstheorie Verw. Geb. **57** (1981a), 223–233.
3. ——————, *Nonparametric estimation of a regression function:limiting distribution*, Austral. J. Statist. **23** (1981b), 186–195.
4. Easterling, R. G., *Discrimination intervals for percentiles in regression*, J. Amer. Statist. Assoc. **64** (1969), 1031–1041.
5. Griffiths, D. and Willcox, M., *Percentile regression:a parametric approach*, J. Amer. Statist. Assoc. **73** (1978), 496–398.
6. Lamperti, J., "Probability," W. A. Benjamin, New York, 1966.

7. Singh, R. S., *On the Glivenko-Cantelli theorem for weighted empiricals based on independent random variables*, Ann. Probability 3 (1975), 371–374.
8. Steinhorst, R. X. and Bowden, D. C., *Discrimination and confidence bands on percentiles*, J. Amer. Statist. Assoc. 66 (1971), 850–854.
9. Stone, C. J., *Consistent nonpapametric regression*, Ann. Statist. 5 (1977), 595–645.
10. Turner, D. L. and Bowden, D. C., *Simulataneous confidence bounds empiricals based on independent random variabvles. Ann. for percentile lines in the general linear model*, J. Amer. Statist. Assoc. 72 (1977), 886–889.
11. Vervaat, W., *Functional central limit theorems for processes with positive drift and their inverses*, Z. Wahrscheinlichkeitstheorie Verw. Geb. 23 (1972), 245–253.