ON NONPARAMETRIC INTERVAL ESTIMATION OF A REGRESSION FUNCTION BASED ON THE RESAMPLING

ALEXANDER ANDRONOV

Riga Technical University

Riga, Latvia

OUTLINE

- 1. Introduction
- 2. Averaging Method
- 3. Median Smoothing
- 4. Case Study
- References

1. INTRODUCTION

We consider nonparametric regression

$$Y = m(x) + \varepsilon \qquad (1.1)$$

where

Y is a dependent variable, $m(\square)$ is an unknown regression function, x is a d-dimensional vector of independent variables (regressors), ε is a random term. It is supposed that the random term has zero expectation ($E(\varepsilon)=0$) and variance $Var(\varepsilon)=\sigma^2\omega(x)$ where σ^2 an unknown constar $\omega(x)$ is a known weighted function. Furthermore we have a sequence of independent observations (Y_i, x_i) $i = 1, 2, .., n$

$\tilde{m}(x)$ $m(x)$ x γ

. On that base we need to construct an upper confidence bound

$$P\{m(x) \le \tilde{m}(x)\} \ge \gamma$$

Usual way [DiCicco and Efron, 1996] consists of using a consistent and asymptotic normal distributed estimate $m(x)$ of $m(x)$. A final expression contains derivatives $m'(x)$, $m''(x)$ and variance σ^2 that are replaced by corresponding estimators.

The resampling approach [Wu, 1986, Andronov and Afanasyeva, 2004] gives an alternative way that can be described as follows. For fixed point x we take k nearest neighbors $x_1^{\bullet}, x_2^{\bullet}, ..., x_k^{\bullet}$ of x among $x_1, x_2, ..., x_n$ in some sense, for example using any kernel function $K_H\left(x - x_i^{\bullet}\right)$ Mahalanobis or other distance):

$$\left\{x_1^{\bullet}, x_2^{\bullet}, ..., x_k^{\bullet}\right\} = \left\{x_i : i \in I_c(x)\right\},$$

where

$$I_c(x) = \left\{i : x_i \text{ is one of the } k \text{ nearest neighbors of } x \text{ among } \left\{x_1, x_2, ..., x_n\right\}\right\}.$$

Now we have sample $(x_1^{\bullet}, Y_1^{\bullet}), (x_2^{\bullet}, Y_2^{\bullet}), \ldots, (x_k^{\bullet}, Y_k^{\bullet})$ instead of $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$.

Then we derive sample without replacement $\{i_1, i_2, \ldots, i_r\}$ of size r ($r < k$) from set $\{1, 2, \ldots, k\}$, form resample $(x_1^{\square}, Y_1^{\square}), (x_2^{\square}, Y_2^{\square}), \ldots, (x_r^{\square}, Y_r^{\square})$, where

$$x_j^{\square} = x_{i_j}^{\bullet} \quad \text{and} \quad Y_j^{\square} = Y_{i_j}^{\bullet},$$

and calculate estimate $m^{\square}(x)$ of our function of interest $m(x)$.

Then we return all selected elements into initial samples and repeat this procedure R times. As result the sequence of estimators

$$m_1^{\square}(x), m_2^{\square}(x), \ldots, m_R^{\square}(x)$$

takes place. After ordering we have sequence

$$m^{\square(1)}(x), m^{\square(2)}(x), \ldots, m^{\square(R)}(x),$$

where $m^{\square(i)}(x) \le m^{\square(i+1)}(x)$.

Let number R is selected so that $R\gamma$ is an integer. Then we set

$$\tilde{m}(x) = m^{\square(R\gamma)}(x).$$

Averaging method and Median smoothing method for $\square m(x)$ ca

are considered. Our main aim is to elaborate a numerical method f probability calculation:

$$\Pr_\gamma(x) = P\{m(x) \le \tilde{m}(x)\}$$

. (1.3)

It means that we need to known a distribution of the

$R\gamma$

$\square m^{(R\gamma)}(x)$

-th order statistic

. That is a main problem that it necessary to solve.

2. AVERAGING METHOD

At first we consider the method of kernel regression estimation [Hardle etc., 2004]. Let $K_H(\square)$ be any kernel function (Epanechnikov, Quartic and so on). Then Nadaraya-Watson point estimator $\overset{\square}{m}(x)$ is calculated by the formula

$$\overset{\square}{m}(x) = \frac{1}{\sum_{i=1}^{r} K_H\left(x - x_i^{\square}\right)} \sum_{i=1}^{r} K_H\left(x - x_i^{\square}\right) Y_i^{\square} \qquad (2.1)$$

where

$x_i^{\square}$ $Y_i^{\square}$

i $i = 1, 2, \ldots, r$

The resampling procedure gives us sequence $\overset{\square}{m}_1(x), \overset{\square}{m}_2(x), \ldots, \overset{\square}{m}_R(x)$

$$\overset{\square}{m}_j(x) = \frac{1}{\sum_{i=1}^{r} K_H\left(x - x_i^{\square}(j)\right)} \sum_{i=1}^{r} K_H\left(x - x_i^{\square}(j)\right) Y_i^{\square}(j) \qquad (2.2)$$

where $x_i^{\square}(j)$ and $Y_i^{\square}(j)$ are the vectors of independent variables and dependent variable for the i-th elements of the j-th, $i = 1, 2, \ldots, r$, $j = 1, 2, \ldots, R$.

With respect to (1.1) we have:

$$E\left(\overset{\square}{m}_j(x) \,\middle|\, x^{\square}(j)\right) = \frac{1}{\sum_{i=1}^{r} K_H\left(x - x_i^{\square}(j)\right)} \sum_{i=1}^{r} K_H\left(x - x_i^{\square}(j)\right) m\left(x_i^{\square}(j)\right)$$

$$Var\left(\hat{m}_j(x) \middle| x^{\square}(j)\right) = \frac{\sigma^2}{\left(\sum_{i=1}^{r} K_H\left(x - x_i^{\square}(j)\right)\right)^2} \sum_{i=1}^{r} \left(K_H\left(x - x_i^{\square}(j)\right)\right)^2 w\left(x_i^{\square}(j)\right)$$

where

$$x^{\square}(j) = \left(x_1^{\square}(j), x_2^{\square}(j), \ldots, x_r^{\square}(j)\right)$$

Then

$$E\left(\hat{m}_j(x)\right) = \frac{1}{\binom{k}{r}} \sum_{z \in \Omega} E\left(\hat{m}_j(x) \middle| z\right) = \frac{1}{\binom{k}{r}} \sum_{z \in \Omega} \left(\frac{1}{\sum_{i=1}^{r} K_H(x - z)} \sum_{i=1}^{r} K_H(x - z) m(z) \right) \quad (2.3)$$

where the sums are taken on Ω - a set of all r-samples without replacement from $\{x_1^{\bullet}, x_2^{\bullet}, \ldots, x_k^{\bullet}\}$.

At first let us calculate the second moment:

$$E\left(\overset{\square}{m}(x)^2\right) = \frac{1}{\binom{k}{r}} E\left(\left(\sum_{z\in\Omega} \overset{\square}{m}(x)\right)^2 \middle| z\right) =$$

$$= \frac{1}{\binom{k}{r}} \sum_{z\in\Omega} \left(\frac{1}{\left(\sum_{i=1}^{r} K_H(x - z_i)\right)^2} E\left(\left(\sum_{i=1}^{r} K_H(x - z_i) Y_i^{\square}(j)\right)^2 \middle| z\right)\right) =$$

$$= \frac{1}{\binom{k}{r}} \sum_{z\in\Omega} \left(\frac{1}{\left(\sum_{i=1}^{r} K_H(x - z_i)\right)^2} \left(\sum_{i=1}^{r} K_H(x - z_i)^2 \left(\sigma^2 w(z_i) + m(z_i)^2\right) + \right.\right.$$

$$\left.\left. + 2\sum_{i=1}^{r-1} \sum_{j=i+1}^{r} K_H(x - z_i) K_H(x - z_j) m(z_i) m(z_j)\right)\right).$$

Now the variance can be calculated by formula

$$Var\left(\tilde{m}(x)\right) = E\left(\tilde{m}(x)\right)^2 - \left(E\left(\tilde{m}(x)\right)\right)^2. \quad (2.4)$$

Now we need to calculate the covariance between two various estimates $\tilde{m}_j(x)$ and $\tilde{m}_{j'}(x)$. We have for $j \neq j'$:

$$Cov(\tilde{m}_j(x), \tilde{m}_{j'}(x)) \overset{d}{=} E\left(\left(\tilde{m}_j(x) - m(x)\right)\left(\tilde{m}_{j'}(x) - m(x)\right)\right) =$$

$$= E\left(\tilde{m}_j(x)\tilde{m}_{j'}(x)\right) - \left(E\left(\tilde{m}(x)\right)\right)^2. \quad (2.5)$$

Further

$$E\left(\hat{m}_j(x)\hat{m}_{j'}(x)\right) = \left(\binom{k}{r}\right)^{-2} \sum_{z\in\Omega}\sum_{v\in\Omega} E\left(\hat{m}_j(x)\hat{m}_{j'}(x) \middle| z, v\right) =$$

$$= \left(E\left(\hat{m}_j(x)\right)\right)^2 + \left(\binom{k}{r}\right)^{-2} \sum_{z\in\Omega} \frac{\sigma^2}{\sum_{i=1}^{r} K_H\left(x - z_i\right)} \times \tag{2.6}$$

$$\times \left(\sum_{v\in\Omega} \frac{1}{\sum_{i=1}^{r} K_H\left(x - v_i\right)} \sum_{z_m \in z \wedge v}^{r} K_H\left(x - z_m\right)^2 w(z_m) \right).$$

Therefore

$$Cov\left(\hat{m}_j(x), \hat{m}_j'(x)\right) = \left(\binom{k}{r}\right)^{-2} \sum_{z\in\Omega} \frac{\sigma^2}{\sum_{i=1}^{r} K_H\left(x - z_i\right)} \times$$

$$\left(\sum_{v\in\Omega} \frac{1}{\sum_{i=1}^{r} K_H\left(x - v_i\right)} \sum_{z_m \in z\wedge v}^{r} K_H\left(x - z_m\right)^2 w(z_m) \right). \quad (2.7)$$

To avoid computational difficulties, it is possible to consider the following estimate instead of (2.1):

$$\hat{m}(x) = \frac{1}{r}\sum_{i=1}^{r} Y_i \hat{}$$

$$\hat{m}_1(x), \hat{m}_2(x), \ldots, \hat{m}_R(x) \quad (2.8)$$

and corresponding those sequence

Lemma 1

Let $Z_1, Z_2, \ldots, Z_k$ be independent random variables with expectations $\mu_1, \mu_2, \ldots, \mu_k$ and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2$. Let $Z_1^{\square}, Z_2^{\square}, \ldots, Z_r^{\square}$ be a random sample of size r from $Z_1, Z_2, \ldots, Z_k$ without replacement and S be their

$$S = Z_1^{\square} + Z_2^{\square} + \ldots + Z_r^{\square}$$

. Then

$$E(S) = \frac{r}{k}\left(\mu_1 + \mu_2 + \ldots + \mu_k\right) \quad (2.9)$$

$$Var(S) = \frac{r}{k}\sum_{j=1}^{k}\left(\sigma_j^2 + \mu_j^2\,\frac{k-r}{k}\right) - 2\,\frac{r(k-r)}{k^2(k-1)}\sum_{j=1}^{k-1}\sum_{i=j+1}^{k}\mu_i\mu_j \quad (2.10)$$