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

Alexander Andronov

Transport and Telecommunication Institute
Lomonosova Str. 1, Riga, LV-1019, Latvia
E-mail: lora@mailbox.riga.lv

A nonparametric regression model \( E(Y) = m(x) \) is considered where \( Y \) is a dependent variable, \( x \) is a \( d \) -dimensional vector of independent variables (regressors) and \( m \) is an unknown function. A sequence of independent observations \( (Y_i, x_i) \), \( i = 1, 2, \ldots, n \), is available. Our aim is to construct an upper confidence bound for \( m(x) \) that corresponds to probability \( \gamma \). The resampling approach is used. The suggested method allows calculating true cover probability.

**Keywords**: nonparametric regression, interval estimation, resampling

1. Introduction

We consider nonparametric regression

\[
Y = m(x) + \varepsilon,
\]

where \( Y \) is a dependent variable, \( m(\cdot) \) is an unknown regression function, \( x \) is a \( d \)-dimensional vector of independent variables (regressors), \( \varepsilon \) is a random term.

It is supposed that the random term has zero expectation \( \mathbb{E}(\varepsilon) = 0 \) and variance \( Var(\varepsilon) = \sigma^2 w(x) \) where \( \sigma^2 \) is an unknown constant and \( w(x) \) is a known weighted function. Furthermore we have a sequence of independent observations \( (Y_i, x_i), \ i = 1, 2, \ldots, n \). On that base we need to construct an upper confidence bound \( \tilde{m}(x) \) for \( m(x) \) at the point \( x \) corresponding to probability \( \gamma \):

\[
P(\tilde{m}(x) \leq m(x)) \geq \gamma.
\]

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

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

\[
\{x_1^*, x_2^*, \ldots, x_k^*\} = \{x_i : i \in I_c(x)\},
\]

where

\[
I_c(x) = \{i : x_i \text{ is one of the } k \text{ nearest neighbours of } x \text{ among } \{x_1, x_2, \ldots, x_n\}\}.
\]

Now we have sample \( (x_1^*, Y_1^*), (x_2^*, Y_2^*), \ldots, (x_k^*, Y_k^*) \) 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 \( (x_{i_1}^*, Y_{i_1}^*), (x_{i_2}^*, Y_{i_2}^*), \ldots, (x_{i_r}^*, Y_{i_r}^*), Y_j^* \), where \( x_{i_j} = x_{i_j}^* \) and \( Y_j^* = Y_{i_j}^* \), and calculate estimate
of our function of interest $m(x)$. Then we return all selected elements into initial samples and we repeat this procedure $R$ times. As a result the sequence of estimators $\hat{m}_1(x), \hat{m}_2(x), ..., \hat{m}_R(x)$ takes place. After ordering we have the sequence $\hat{m}^{(1)}(x), \hat{m}^{(2)}(x), ..., \hat{m}^{(R)}(x)$, where $\hat{m}^{(1)}(x) \leq \hat{m}^{(j)}(x)$.

Let the number $R$ is selected so that $R\gamma$ is an integer. Then we set $\hat{m}(x) = \hat{m}^{(R\gamma)}(x)$.

In the presented paper averaging method of estimator $\hat{m}(x)$ forming is considered. Our main aim is to elaborate a numerical method for cover probability calculation:

$$\Pr_x \{m(x) \leq \hat{m}(x)\}.$$

(1.3)

It means that we have to know a distribution of the $R\gamma$-th order statistics $\hat{m}^{(R\gamma)}(x)$. This is a main problem that is necessary to be solved.

2. Averaging Method

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

$$\hat{m}(x) = \frac{1}{\sum_{i=1}^{r} K_H(x-x_i^o)} \sum_{i=1}^{r} K_H(x-x_i^o)Y_i^o,$$

(2.1)

where $x_i^o$ and $Y_i^o$ are a vector of independent variables and dependent variable for the $i$-th elements of the resample, $i = 1, 2, ..., r$.

The resampling procedure gives us sequence $\hat{m}_1(x), \hat{m}_2(x), ..., \hat{m}_R(x)$,

$$\hat{m}_j(x) = \frac{1}{\sum_{i=1}^{r} K_H(x-x_i^j)} \sum_{i=1}^{r} K_H(x-x_i^j)Y_i^j,$$

(2.2)

where $x_i^j$ and $Y_i^j$ are a vector of independent variables and dependent variable for the $i$-th elements of the $j$-th resample, $i = 1, 2, ..., r, j = 1, 2, ..., R$.

With respect to (1.1) we have:

$$E(\hat{m}(x)|x^o(j)) = \frac{1}{\sum_{i=1}^{r} K_H(x-x_i^o(j))} \sum_{i=1}^{r} K_H(x-x_i^o(j))m(x_i^o(j)),$$

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

where $x^o(j) = (x_1^o(j), x_2^o(j), ..., x_r^o(j))$.

Then

$$E(\hat{m}(x)) = \frac{1}{k} \sum_{z \in \Omega} E(\hat{m}(x)|x^o) = \frac{1}{k} \sum_{z \in \Omega} \left( \frac{1}{r} \sum_{i=1}^{r} K_H(x-z_i) m(z_i) \right),$$

(2.3)
where the sums are taken on set $\Omega$ of all $r$-samples $z = (z_1, z_2, \ldots, z_r)$ without replacement from the set $\{x_1^*, x_2^*, \ldots, x_k^*\}$.

Analogous expression we can to write down for unconditional variance. At first let us calculate the second moment:

$$E(\hat{m}(x)^2) = \frac{1}{k} \sum_{r \in \Omega} \frac{1}{r} \left\{ \sum_{i=1}^{k} \frac{1}{K_H(x-z_i)} E\left( \left( \sum_{j=1}^{r} K_H(x-z_j) \right)^2 \right) \right\} =$$

$$= \frac{1}{k} \sum_{r \in \Omega} \frac{1}{r} \left\{ \sum_{i=1}^{k} \frac{1}{K_H(x-z_i)} \left( \sum_{i=1}^{r} \frac{1}{K_H(x-z_i)} \left( \sum_{j=1}^{r} \frac{1}{K_H(x-z_j)} \right)^2 \right) \right\}.$$  

Now the variance can be calculated by the following formula

$$Var(\hat{m}(x)) = E(\hat{m}(x)^2) - (E(\hat{m}(x)))^2.$$  

(2.4)

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

$$Cov(\hat{m}_j(x), \hat{m}_{j'}(x)) = E\left( (\hat{m}_j(x) - m(x))(\hat{m}_{j'}(x) - m(x)) \right) = E(\hat{m}_j(x)\hat{m}_{j'}(x)) - (E(\hat{m}(x)))^2.$$  

(2.5)

Further

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

$$= \left( E(\hat{m}_j(x)) \right)^2 + \left( \frac{k}{r} \right)^2 \sum_{z \in \Omega} \sum_{v \in \Omega} \sigma^2 \frac{1}{K_H(x-z_i)} \left( \frac{1}{K_H(x-v_i)} \sum_{z \in \Omega} \sum_{z' \in \Omega} K_H(x-z) K_H(x-z') \right) \frac{w(z)}{w(z')}.$$  

(2.6)

Therefore

$$Cov(\hat{m}(x)) = \left( \frac{k}{r} \right)^2 \sum_{z \in \Omega} \sum_{v \in \Omega} \sigma^2 \frac{1}{K_H(x-z_i)} \left( \frac{1}{K_H(x-v_i)} \sum_{z \in \Omega} \sum_{z' \in \Omega} K_H(x-z) K_H(x-z') \right) \frac{w(z)}{w(z')}.$$  

(2.7)

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

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

(2.8)

and the corresponding sequence $\hat{m}_1(x), \hat{m}_2(x), \ldots, \hat{m}_k(x)$.

Expectations, variances and covariance matrix for this sequence of random variables can be determined using the following lemmas.
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 \( Z_1^*, Z_2^*, \ldots, Z_r^* \) be a random sample of size \( r \) from \( Z_1, Z_2, \ldots, Z_k \) without replacement and \( S \) be their sum: \( S = Z_1^* + Z_2^* + \ldots + Z_r^* \). Then

\[
E(S) = \frac{r}{k} (\mu_1 + \mu_2 + \ldots + \mu_k),
\]

(2.9)

\[
Var(S) = \frac{1}{r k} \sum_{j=1}^{k} \left( \sigma_j^2 + \mu_j^2 \frac{k-r}{k} \right) - 2 \frac{r(k-r)}{k^2(k-1)} \sum_{j=1}^{k-1} \sum_{i=j+1}^{k} \mu_i \mu_j.
\]

(2.10)

Lemma 2.
For the conditions of the previous Lemma let the sample \( Z_1^*, Z_2^*, \ldots, Z_r^* \) be returned into the set \( \{ Z_1^*, Z_2^*, \ldots, Z_k^* \} \) and the described procedure be repeated, so that we have new sample \( Z_1^{*'}, Z_2^{*'}, \ldots, Z_r^{*'} \) and a corresponding sum \( S^{*'} = Z_1^{*'} + Z_2^{*'} + \ldots + Z_r^{*'} \). Then the covariance between \( S \) and \( S^{*'} \) is calculated by the formula

\[
Cov(S, S^{*'}) = \left( \frac{r}{k} \right)^2 \sum_{i=1}^{k} \sigma_i^2.
\]

(2.11)

In our case \( Y_i^{*'} \) and \( Y_i^{*'} \) play the part of \( Z_i \) and \( Z_i^* \) correspondingly, \( \bar{m}(x) \) is equal to \( S/r \). Furthermore \( \mu_i = \bar{m}(x_i) \) and instead of \( \sigma_i^2 \) must be \( \sigma^2 \bar{m}(x_i) \).

With respect to the given suppositions, random vector \( (\bar{m}_1(x), \bar{m}_2(x), \ldots, \bar{m}_r(x)) \) has multi-dimensional symmetric distribution with characteristics determined by (2.3), (2.4), (2.7) or (2.9)-(2.11). Therefore to calculate coverage probability (1.3) means to calculate the probability that at last \( R(1 - \gamma) \) components of vector \( (\bar{m}_1(x), \bar{m}_2(x), \ldots, \bar{m}_r(x)) \) will be greater than \( m(x) \). For this it is possible to use normal approximation of the distribution. Unfortunately again we are faced with a hard computational problem. Usually for that solving crude Monte Carlo method is used.

APPENDIX

Proof of Lemma 1.
Let \( \chi_j = 1 \) if the random variable \( Z_j \) belongs to the sample \( \{ Z_1^*, Z_2^*, \ldots, Z_r^* \} \) and \( \chi_j = 0 \) otherwise. Of course \( \chi_1, \chi_2, \ldots, \chi_k \) are dependent random variables because \( \chi_1 + \chi_2 + \ldots + \chi_k = r \). We have: \( P(\chi_j = 1) = r/k, P(\chi_j = 0) = 1 - r/k, E(\chi_j) = P(\chi_j = 1) = r/k \), \( Var(\chi_j) = (1-r/k) r/k, E(\chi_i \chi_j) = P(\chi_i = 1, \chi_j = 1) = r(r-1) / (k(k-1)) \) for \( i \neq j \). Furthermore

\[
S = \sum_{i=0}^{k} \chi_i Z_i.
\]

Random variables \( \chi_i \) and \( Z_i \) are independent therefore

\[
E(S) = \sum_{i=1}^{k} E(\chi_i Z_i) = \sum_{i=1}^{k} E(\chi_i) E(Z_i) = \frac{r}{k} \sum_{i=1}^{k} \mu_i,
\]

\[
E((\chi_i Z_i)^2) = E(\chi_i^2) E(Z_i^2) = \frac{r}{k} (\mu_i^2 + \sigma_i^2),
\]
Part II. Statistical Inferences

\[ \text{Var}(\chi_i Z_i) = E((\chi_i Z_i)^2) - (E(\chi_i Z_i))^2 = \]

\[ = \frac{r}{k} (\mu_i^2 + \sigma_i^2) - \left( \frac{r}{k} \mu_i \right)^2 = \frac{r}{k} \left( \sigma_i^2 + \mu_i^2 \left( 1 - \frac{r}{k} \right) \right). \tag{A.1} \]

Random variables \( Z_i, Z_j \) and \( \chi_i, \chi_j \) for \( i \neq j \) are independent, too, therefore

\[ E(\chi_i Z_i, \chi_j Z_j) = E(\chi_i, \chi_j) E(Z_i) E(Z_j) = \mu_i \mu_j \frac{r(r-1)}{k(k-1)}, \]

\[ \text{Cov}(\chi_i Z_i, \chi_j Z_j) = \mu_i \mu_j \frac{r(r-1)}{k(k-1)} - \mu_i \mu_j \left( \frac{r}{k} \right)^2 = -\mu_i \mu_j \frac{r(k-r)}{k^2(k-1)}. \tag{A.2} \]

Formulas (A.1) and (A.2) give formula (2.10).

**Proof of Lemma 2.**

Let \( S = \sum_{i=1}^{k} \chi_i Z_i, \ S^* = \sum_{j=1}^{k} \chi_j^* Z_j. \)

Then

\[ \text{Cov}(S, S^*) = \text{Cov} \left( \sum_{i=1}^{k} \chi_i Z_i, \sum_{j=1}^{k} \chi_j^* Z_j \right) = \sum_{i=1}^{k} \sum_{j=1}^{k} \text{Cov}(\chi_i Z_i, \chi_j^* Z_j) = \]

\[ = \sum_{i=1}^{k} \sum_{j=1}^{k} \text{Cov}(\chi_i Z_i, \chi_j^* Z_j) + \sum_{i=1}^{k} \sum_{j=1}^{k} \text{Cov}(\chi_i Z_i, \chi_j^* Z_j) \]

For \( i \neq j \) random variables \( \chi_i, \chi_j^*, Z_i, Z_j \) are independent, therefore \( \text{Cov}(\chi_i Z_i, \chi_j^* Z_j) = 0. \) Further

\[ \text{Cov}(\chi_i Z_i, \chi_i^* Z_i) = E(\chi_i^* Z_i^2) - E(\chi_i Z_i) E(\chi_i^* Z_i) = \]

\[ = E(\chi_i) E(\chi_i^* Z_i) E(Z_i^2) - \left( \frac{r}{k} \mu_i \right)^2 = \left( \frac{r}{k} \sigma_i \right)^2. \]

Therefore

\[ \text{Cov}(S, S^*) = \left( \frac{r}{k} \right)^2 \sum_{i=1}^{k} \sigma_i^2. \]

**References**