Computer Modelling and New Technologies, 2007, Vol.11, No.1, 57-61
Transport and Telecommunication Institute, Lomonosova 1, Riga, LV-1019, Latvia

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 d – a 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 γ. 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

$$Y = m(x) + \varepsilon , \tag{1.1}$$

where Y is a dependent variable, $m(\circ)$ 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 $Var(\varepsilon) = \sigma^2 w(x)$ where σ^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), $i = 1, 2, .., n$. On that base we need to construct an upper confidence bound $\widetilde{m}(x)$ for $m(x)$ at the point x corresponding to probability γ:

$$P\{m(x) \le \widetilde{m}(x)\} \ge \gamma . \tag{1.2}$$

Usual way [DiCicco and Efron, 1996] consists of using a consistent and asymptotic normal distributed estimate $\widehat{m}(x)$ of $m(x)$. A final expression contains derivatives $m'(x)$, $m''(x)$ and variance σ^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 $x_1^\bullet, x_2^\bullet, \ldots, x_k^\bullet$ of x among $x_1, x_2, \ldots, x_n$ (in some sense, for example using any kernel function $K_H(x - x_i^\bullet)$, Mahalanobis or other distance):

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

where

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

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 $(x_1^\circ, Y_1^\circ), (x_2^\circ, Y_2^\circ), \ldots, (x_r^\circ, Y^\circ)$, where $x_j^\circ = x_{i_j}^\bullet$ and $Y_j^\circ = Y_{i_j}^\bullet$, and calculate estimate

$\widehat{m}(x)$ 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 $\widehat{m}_1(x), \widehat{m}_2(x), ..., \widehat{m}_R(x)$ takes place. After ordering we have the sequence $\widehat{m}^{(1)}(x), \widehat{m}^{(2)}(x), ..., \widehat{m}^{(R)}(x)$, where $\widehat{m}^{(i)}(x) \le \widehat{m}^{(i+1)}(x)$.

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

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

$$\Pr_\gamma(x) = P\{m(x) \le \widetilde{m}(x)\}. \tag{1.3}$$

It means that we have to know a distribution of the $R\gamma$ -th order statistics $\widehat{m}^{(R\gamma)}(x)$. 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 $K_H(\circ)$ be any kernel function (Epanechnikov, Quartic and so on). Then Nadaraya-Watson estimator $\widehat{m}(x)$ is calculated by the following formula

$$\widehat{m}(x) = \frac{1}{\sum_{i=1}^{r} K_H\left(x - x_i^\circ\right)} \sum_{i=1}^{r} K_H\left(x - x_i^\circ\right) Y_i^\circ, \tag{2.1}$$

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

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

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

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

With respect to (1.1) we have:

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

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

where $x^\circ(j) = \left(x_1^\circ(j), x_2^\circ(j), ..., x_r^\circ(j)\right)$.

Then

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

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

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

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

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

Now the variance can be calculated by the following formula

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

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

$$Cov(\hat m_j(x),\hat m_{j'}(x))\overset{d}{=}E\left((\hat m_j(x)-m(x))(\hat m_{j'}(x)-m(x))\right)=E\left(\hat m_j(x)\hat m_{j'}(x)\right)-\left(E(\hat m(x))\right)^2. \tag{2.5}$$

Further

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

$$=\left(E(\hat m_j(x))\right)^2+\left(\binom{k}{r}\right)^{-2}\sum_{z\in\Omega}\frac{\sigma^2}{\sum_{i=1}^{r}K_H(x-z_i)}\left(\sum_{v\in\Omega}\frac{1}{\sum_{i=1}^{r}K_H(x-v_i)}\sum_{z_m\in z\wedge v}^{r}K_H(x-z_m)^2w(z_m)\right). \tag{2.6}$$

Therefore

$$Cov(\hat m(x))=\left(\binom{k}{r}\right)^{-2}\sum_{z\in\Omega}\frac{\sigma^2}{\sum_{i=1}^{r}K_H(x-z_i)}\left(\sum_{v\in\Omega}\frac{1}{\sum_{i=1}^{r}K_H(x-v_i)}\sum_{z_m\in z\wedge v}^{r}K_H(x-z_m)^2w(z_m)\right). \tag{2.7}$$

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

$$\hat m(x)=\frac{1}{r}\sum_{i=1}^{r}Y_i^\circ \tag{2.8}$$

and the corresponding sequence $\hat m_1(x),\hat m_2(x),...,\hat m_R(x)$.

Expectations, variances and covariance matrix for this sequence of random variables can be determined using the following lemmas.

Lemma 1.

Let $Z_1, Z_2, ..., Z_k$ be independent random variables with expectations $\mu_1, \mu_2, ..., \mu_k$ and variances $\sigma_1^2, \sigma_2^2, ..., \sigma_k^2$. Let $Z_1^\circ, Z_2^\circ, ..., Z_r^\circ$ be a random sample of size *r* from $Z_1, Z_2, ..., Z_k$ without replacement and *S* be their sum: $S = Z_1^\circ + Z_2^\circ + ... + Z_r^\circ$. Then

$$E(S) = \frac{r}{k}\left(\mu_1 + \mu_2 + ... + \mu_k\right), \tag{2.9}$$

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

Lemma 2.

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

$$Cov\left(S, S^\bullet\right) = \left(\frac{r}{k}\right)^2\sum_{i=1}^{k}\sigma_i^2. \tag{2.11}$$

In our case $Y_i^\bullet$ and Y_i° play the part of Z_i and Z_i° correspondingly, $\widehat{m}(x)$ is equal to *S*/*r*. Furthermore $\mu_i = m\left(x_i^\bullet\right)$ and instead of σ_i^2 must be $\sigma^2 w\left(x_i^\bullet\right)$.

With respect to the given suppositions, random vector $\left(\widehat{m}_1(x), \widehat{m}_2(x), ..., \widehat{m}_R(x)\right)$ has multi-dimensional symmetric distribution with characteristics determined by (2.3), (2.4), (2.7) or (2.9)-(2.11). Therefore to calculate cover probability (1.3) means to calculate the probability that at last $R(1 - \gamma)$ components of vector $\left(\widehat{m}_1(x), \widehat{m}_2(x), ..., \widehat{m}_R(x)\right)$ 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 $\{Z_1^\circ, Z_2^\circ, ..., Z_r^\circ\}$ and $\chi_j = 0$ otherwise. Of course $\chi_1, \chi_2, ..., \chi_k$ are dependent random variables because $\chi_1 + \chi_2 + ... + \chi_k = r$. 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 χ_i and Z_i are independent therefore

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

$$E\left((\chi_i Z_i)^2\right) = E(\chi_i^2)E(Z_i^2) = \frac{r}{k}(\mu_i^2 + \sigma_i^2),$$

$$Var(\chi_i Z_i) = E\left((\chi_i Z_i)^2\right) - \left(E(\chi_i Z_i)\right)^2 =$$

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

Random variables Z_i, Z_j and χ_i χ_j for $i \neq j$ are independent, too, therefore

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

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

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

Proof of Lemma 2.
Let

$$S = \sum_{i=1}^{k} \chi_i Z_i, \; S^\bullet = \sum_{j=1}^{k} \chi_j^\bullet Z_j.$$

Then

$$Cov(S, S^\bullet) = Cov\left(\sum_{i=1}^{k} \chi_i Z_i, \sum_{j=1}^{k} \chi_j^\bullet Z_j\right) = \sum_{i=1}^{k}\sum_{j=1}^{k} Cov\left(\chi_i Z_i, \chi_j^\bullet Z_j\right) =$$

$$= \sum_{i=1}^{k} Cov\left(\chi_i Z_i, \chi_i^\bullet Z_i\right) + \sum_{i=1}^{k}\sum_{j \neq i} Cov\left(\chi_i Z_i, \chi_j^\bullet Z_j\right).$$

For $i \neq j$ random variables $\chi_i, \chi_j^\bullet, Z_i, Z_j$ are independent, therefore $Cov\left(\chi_i Z_i, \chi_j^\bullet Z_j\right) = 0$. Further

$$Cov\left(\chi_i Z_i, \chi_i^\bullet Z_i\right) = E\left(\chi_i \chi_i^\bullet Z_i^2\right) - E\left(\chi_i Z_i\right)E\left(\chi_i^\bullet Z_i\right) =$$

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

Therefore

$$Cov\left(S, S^\bullet\right) = \left(\frac{r}{k}\right)^2 \sum_{i=1}^{k} \sigma_i^2.$$

References

1. Andronov, A., Afanasyeva, H. Resampling-based nonparametric statistical inferences about the distributions of order statistics. In: *Transactions of XXIV International Seminar on Stability Problems for Stochastic Models*. Riga: Transport and Telecommunication Institute, 2004, pp. 300-307.
2. DiCicco, T. J., Efron, B. Bootstrap confidence intervals, *Statistical Sciences*, Vol. 11, No 3, 1996, pp. 189-228.
3. Hardle, W., Muller, M., Sperlich, S., Werwatz, A. *Nonparametric and Semiparametric Models*. Berlin: Springer, 2004.
4. Wu, C.F.J. Jackknife, Bootstrap and other resampling methods in regression analysis, *The Annals of Statistics*, Vol. 14, No 3, 1986, pp. 1261-1295.