Bandwidth Selection Problem for Nonparametric Regression Model with Right-Censored Data

Dursun AYDIN (daydin@mu.edu.tr)
Ersin YILMAZ (ersinyilmaz@mu.edu.tr)
Mugla Sitki Kocman University, Turkey

ABSTRACT

In this paper, the proposed estimator for the unknown nonparametric regression function is a Nadarya-Watson (Nadarya, 1964; Watson, 1964) type kernel estimator. In this estimation procedure, the censored observations are replaced by synthetic data points based on Kaplan-Meier estimator. As known performance of the kernel estimator depends on the selection of a bandwidth parameter. To get an optimum parameter we have considered six selection methods such as the improved version of Akaike information criterion (AICc), Bayesian information criterion (BIC), generalized cross validation (GCV), risk estimation with classical pilots (RECP), Mallow's Cp criterion and restricted empirical likelihood (REML), respectively. In addition, we discuss the behavior of the estimators obtained by these selection methods under different configurations of the censoring level and sample sizes. Simulation and real lifetime data results are presented to evaluate and compare the performance of the selection methods. Thus, a optimum criterion is provided for smoothing parameter selection.

Key-Words: *Kernel Smoothing, Kaplan-Meier Estimator, Nonparametric Regression, Censored data*

1.INTRODUCTION

Censored data arises in a number of applied fields, such as medical, sociology, biology, risk theory, demography, and other appropriate areas. Observations in these fields are usually incomplete, especially in medical studies. For example, some patients may still be alive, disease-free or die at the termination of a medical study. Hence, rather than an observation of a patient's lifetime we observe only the minimum of the lifetime and a censoring time.

There are mainly two conventional statistical methods used in analysis of the functional relationship between covariates and censored response, known as lifetime or failure time. One of these methods is parametric, if the distribution of lifetime is known. The second is nonparametric, if distribution of lifetime is unknown. Although the parametric methods can be simple and efficient if the model is correctly specified, they are not widely used in general, since their restrictions and assumptions on the model. Instead, we focus on the nonparametric methods do not require the knowledge of the underlying distribution of the lifetime.

Let's consider the nonparametric regression model given by

$$Y_i = g(X_i) + \varepsilon_i \qquad (1)$$

where Y_i's are response values and X_i's are the values of the explanatory variable and ε_i's are independent and identically distributed random errors with zero mean and constant variance σ^2 and g is an unknown function.

In our study, we are interested in estimating the unknown function $g(.)$ when Y is observed incompletely and right censored by a random variable *C*, but T_i's are observed completely. Therefore, instead of observing (Y_i, X_i) we observe $\{(X_i, T_i, \delta_i), i \leq 1 \leq n\}$ with

$$T_i = \min(Y_i, C_i)_{i=1}^{n} \text{ with } \delta_i = I(Y_i \leq C_i) \qquad (2)$$

where T_i's are the observations of the updated response variable with unknown distribution *K* and $\delta_i = \mathrm{I}(.)$ is the sign function that indicates the existence of the censorship. If there is a censorship on response variable then and otherwise . In order to provide the consistency and accuracy of the model (1), we need some assumptions for distribution of (*X*, *Y*, δ) such that

i. *Y* and *C* are independent and unknown distributed as *F* and *G*, respectively. Also, *F* and *G* have no jumps in common

ii. $\mathrm{P(Y \leq C \mid Y, X) = P(Y \leq C \mid Y)}$

The first assumption is the common censorship assumption when we estimate the right censored data. The jump assumption does not exclude discontinuities of *F* and *G* at distinct points. The second assumption means that given response variable, we cannot obtain any more information from the covariate whether there is a censorship or not. See Stute (1993) for additional details on the second assumption.

In this paper we propose a Nadarya-Watson kernel type smoothing to fit model (1) when response variable T is at risk of being censored. Efficient implementation of this smoothing method requires a proper smoothing parameter. The mentioned parameter is determined by the selection methods, such as AICc, BIC, GCV, Cp, RECP, and REML, respectively. In this context, this paper basically presents and compares these estimates of the lifetime T given the covariate X under censorship.

Many authors have dealt with the estimation problem of the nonparametric regression model based on kernel smoothing. Examples of this work include Watson (1964), Wong (1983), Vieu (1991), Terrel and Scott (1992), Hardle (1990), Stute (1993), and Hardle et. al., (1997). Also, a number of authors consider the kernel smoothing for estimating the nonparametric function based on censored data. For example, Kaplan and Meier's (1958) product limit method is the most commonly used technique for estimating the survival function. Koul et. al. (1981) proposed the synthetic data generation for estimation of right-censored data. Leurgans (1987) studied random censoring and synthetic data for linear models. Zheng (1984) made a dissertation about regression with censored data. Recently, empirical likelihood semi-parametric random censorship models are discussed by Wang and Li (2002).

According to organization of this paper, fundamental ideas are examined in section 2. In Section 3, the kernel type estimators in nonparametric models are discussed. Estimating risk and efficiency are examined in Section 4. Section 5 reviews six different bandwidth selection methods. Section 6 compares these methods via simulated data sets. In Section 7, a real data example is given. Finally, the concluding remarks are presented in section 8. Proofs and supplemental technical materials are relegated to the Appendix.

2. THE FUNDAMENTAL IDEAS

Let $T_1,...,T_n$ and $\delta_1,...,\delta_n$ be nonnegative independent and identically distributed random variables. In the model (2), T_i 's are the observed lifetimes, while δ_i's store up the information whether an observation is censored or uncensored. Moreover, we denote the unknown probability distribution functions of the lifetimes, the censoring times, and the observed lifetimes as $F(z) = P(Y > z)$, $G(z) = P(C > z)$, and $K(z) = P(T > z)$, $(z \in R)$, respectively. Also, for these probabilities it can be defined as three supremum points

$$\upsilon_F = \sup\{z \in R : F(z) > 0\},$$

$$\upsilon_G = \sup\{z \in R : G(z) > 0\},$$

and

$$\upsilon_K = \sup\{z \in R : K(z) > 0\}.$$

Because of the independence of *Y* and *C*, the unknown distribution function of observed lifetimes can be written as

$$K(z) = F(z).G(z) = P(T > z)$$

In order to ensure that model is identifiable, we assume that

$$\upsilon_K = \sup\{z \in R : K(z) > 0\} = \min\{\upsilon_F, \upsilon_G\}$$

One of the main goals in this paper is also to estimate the unknown distribution functions *F* and *G*. If we use uncensored data, the nonparametric estimate of function *F* can be obtained by

$$\hat{F}_n(z)=\hat{F}(z,\{Y_1,...,Y_n\})=\frac{1}{n}\sum_{i=1}^{n}\mathrm{I}_{[Y_i>z]} \tag{3}$$

where $\mathrm{I}_{Y_i>z}$ is an indicator function for the values of lifetime. It is well known that the Glivenko-Cantelli theorem extends the law of large numbers and gives the uniform convergence. This theorem implies that $\hat{F}_n(z)$ is a strongly uniform consistent estimate *F*(z). In other words, uniform convergence is given by (see Van der Vaart, 1998¸ Stute and Wang 1993)

$$\left\|\hat{F}_n-F\right\|_\infty=\sup_{t\in R}\left|\hat{F}_n(z)-F(z)\right|\to 0 \tag{4}$$

Because of the censorship, $\hat{F}_n(z)$ cannot be directly calculated by equation (3). The most important reason for this case, the number of lifetime greater than *z* are not exactly known for all $z\in[0,\upsilon_F]$. In this case, it is needed to fined a nonparametric estimate of *F*. It is emphasized that the unknown estimate of *F* can be provided by Kaplan-Meier estimator (Kaplan and Meier, 1958).

$$\hat{F}_n(z)=\prod_{\substack{t=1\\T_{(i)}\le z}}^{n}\left[\frac{n-i}{n-i+1}\right]^{\delta_{(i)}}\quad(z\in R),\, i=1,...,n \tag{5}$$

The estimates for distribution *G* are similar to those for *F* in (5), given by

$$\hat{G}_n(z)=\prod_{\substack{t=1\\T_{(i)}\le z}}^{n}\left[\frac{n-i}{n-i+1}\right]^{1-\delta_{(i)}}\quad(z\in R),\, i=1,...,n \tag{6}$$

Where $T_{(1)}\le ...\le T_{(n)}$ are the ordered values of observed lifetimes and $\delta_{(1)}\le ...\le\delta_{(n)}$ are the corresponding censoring indicators connected with observed lifetimes T_i. It is also note that ties among lifetimes and censoring times are defined by

If $T_{(i)}=T_{(j)},\, \delta_{(i)}=\delta_{(j)}=1\Rightarrow i<j$ (7)

As in explained above, in this ordering censored lifetime data points ($\delta_{(i)}=0$) take place before uncensored data points ($\delta_{(i)}=1$). Moreover, we have $0\le T_{(1)}\le ...\le T_{(n)}\le\upsilon_K$ where, $T_{(n)}$ is the largest value of the ordered sequence. In this case, as $n\to\infty$, $T_{(n)}\to\upsilon_K$ (see Peterson, 1977).

3. ESTIMATION METHOD

Let (X,Y) be random variables vector taking values in $R^d\times R$. In nonparametric regression model (1) we want to find an estimation of the function $g(X)=E\left(Y\,\middle|X=X\right),(X\in R^d)$ from the data

$\{(X_1, Y_1),...,(X_n, Y_n)\}$

Suppose that we are interesting a class of estimators for g, $\mathrm{B(H)} = [g_h : h \in \mathrm{H}]$, with H denoting any index (known as bandwidth or smoothing parameter) set. The mention index parameter h may be any scalar or vector. Also, let $g_h(X)$ be a function based on bandwidth h. The L_2 loss in estimating g is defined by

$$L_2(h) = n^{-1}\sum_{i=1}^{n}(g_i - g_{hi}(X))^2 \tag{8}$$

where g_{hi} is i.th entries of the n-vector g_h. Note that this squared Euclidean distance (8) between g and g_h measures the closeness of g_h to g. The expected value of the L_2 loss is so-called R_2 risk, given by

$$R_2(h) = E\left\{n^{-1}\sum_{i=1}^{n}(g_i - g_{hi}(X))^2\right\} \tag{9}$$

In here, the key idea is to find a regression function g with $g_h(X)$ close to $g(X)$. Such a regression function minimizes the R_2 risk over all measurable functions $g : R^d \to R$.

Another measure that is connected to (16) is the P_2 risk, sometimes called as mean square error (MSE) of prediction. The P_2 risk is

$$P_2(h) = \sigma^2 + R_2(h) \tag{10}$$

where σ^2 is a variance of random error terms (see Eubank, 1988).

Since the estimators in B(H) are obtained by elements in index set H, an optimum estimator can be described with an index value h minimizes the R_2 or P_2 risk. But these risk measures cannot be computed directly because of they depend on unknown true regression function g and a smoothing parameter h.

Let's consider the equation (2). In the regression analysis, one wants to estimate T from the data

$\{(X_1, T_1, \delta_1),...,(X_1, T_1, \delta_1)\}$

The conditional expected value of the regression function at a point X can be obtained by averaging those T_i's where X_i is close to X. Such an estimate can be obtained by kernel smoothing. Because of the censoring mechanism, for estimating $g(.)$ ordinary kernel smoothing method can not be applied directly here. To overcome this problem we considered the new response observations in (2). Also, we transformed the right-censored variable "*T*" into synthetic variable "'$T_{\hat{G}}$" (see Koul et. al.,1981). In practice, because of the values *T* are censored observations, the censoring distribution *G* is usually unknown. In this case, Koul et al. (1981) proposed to replace *G* by its Kaplan-Meier estimator in (6).

Using the equation (6) the synthetic response variable can be obtained as

$$T_{i\hat{G}} = \left(\delta_i T_i\right) \Big/ \left(1-\hat{G}(T_i)\right), \; i = 1,2,\ldots,n \qquad (11)$$

From this synthetic data, the model (1) can be rewritten as

$$\mathbf{T}_{\hat{G}} = \left\{\mathbf{g} = \left(g\left(X_1\right)+\ldots+g\left(X_n\right)\right)\right\} + \boldsymbol{\varepsilon} \qquad (12)$$

where $\boldsymbol{\varepsilon} = \mathbf{T}_{\hat{G}} - \mathbf{g}$. Conceptually, as $n \to \infty$, $\mathrm{E}\left(\boldsymbol{\varepsilon}\right) \cong 0$. This information will help us to define estimates for the function in (12). Then, kernel smoothing can be used as a nonparametric approcah to get a proper estimate of the $g(.)$ in (2).

The kernel smoothing is one of the most widely used methods, which considers a weighted average of the data. Let $\hat{T}_{i\hat{G}}$ be a kernel smoother estimate of the i.th response observation. Then, a kernel smoother is defined as follows

$$\hat{T}_{i\hat{G}h} = \sum_{j=1}^{n} w_{ij} t_j \qquad (13)$$

where t_j's are elements of the synthetic response variable $\hat{T}_{i\hat{G}}$ and w_{ij}'s are known as kernel weights given by Nadaraya-Watson (1964). The specific weights for the kernel smoothing is expressed as

$$W_{ij} = \frac{K\left(\dfrac{x-X_j}{h}\right)}{\sum_{j=1}^{n} K\left(\dfrac{x-X_j}{h}\right)} = \frac{K(u)}{\sum K(u)} \qquad (14)$$

where h is called bandwidth, and $\sum w_{ij} = 1$. The function $K(u)$ determines the shape of the regression curves, while the parameter h determines their width and also governs the amount of averaging.

It comes out that the kernel estimator expressed in (13) is a weighted average of the response with right censored data. This approach is a called kernel smoothing because of a kernel function, *K*, to determine the weights. These kernel functions have the following properties:

- $K\left(u\right) \geq 0$ for all u,
- $K\left(u\right) = K\left(u-\right) = K\left(u\right)$ and
- $\int K\left(u\right) du = 1$.

For example, Gaussian kernel function,

$$K_G(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right), u \in [-\infty, +\infty]$$

and other alternative kernel functions provide the properties of the kernel weight function, $K\left(u\right)$. Also, in order to ensure that kernel estimator is consistent, we assume that

If $\left|u\right| \to \infty$, and $E\left(T^2\right) < \infty$ then $\int \left|K(u)\right| du < \infty, \quad u K(u) \to 0$ and suppose that $h \to 0, \; nh \to \infty$ then it can be seen that

$$\frac{1}{n}\sum_{i=1}^{n} W_{ij}t_{ij} = \hat{g}(X_i) \xrightarrow{P} g(X_i) \tag{15}$$

where $\xrightarrow{P}$ denotes "convergence in probability" according to Slutsky's theorem (1925).

The kernel smoother (13) also can be rewritten as in matrix form

$$\hat{\mathbf{T}}_{\hat{G}h} = \mathbf{W}_h \mathbf{T}_{\hat{G}} = \hat{\mathbf{g}}_h \tag{16}$$

where $\mathbf{W}_h = \left[w_{ij} \right]$ is a kernel smoother matrix based on the parameter h.

4. ESTIMATING THE RISK AND EFFICIENCY

In previous section the L_2 loss, R_2 and P_2 risks are considered as a measure of performance of an estimator of g. Here we will focus on the estimating of these risks measures for kernel estimator using right censored data. According to $\mathrm{B}(\mathrm{H}) = \left[g_h : h \in \mathrm{H} \right]$, it can be said that for each h there is an $n \times n$ smoother matrix $\mathbf{W}_h$ in (16). Accordingly, the equation (13) can be rewritten as

$$\hat{\mathbf{g}}_h = \left(\hat{g}_h(X_1), ..., \hat{g}_h(X_n) \right)' = \mathbf{W}_h \mathbf{T}_{\hat{G}} \tag{17}$$

Also, it is assumed that $\mathbf{W}_h$ is a positive semi-definite and symmetrical matrix.

The main goal is to select an appropriate estimator of g from among the elements $\left[\hat{g}_h : h \in \mathrm{H} \right]$. In order to find an optimum estimator there are some performance measures which are widely used and accepted. The P_2 risk in (10), one of these measures can be obtained by average value of residuals sum of squares $n^{-1}RSS(h)$. The mentioned residual sum of squares (*RSS*) is defined as

$$RSS(h) = \sum_{i=1}^{n} \left((\hat{g}_h)_i - T_{i\hat{G}} \right)^2 \tag{18}$$

In matrix form, equation (18) can be stated as

$$\begin{aligned} RSS(h) &= \left(\hat{\mathbf{g}}_h - \mathbf{T}_{\hat{G}} \right)' \left(\hat{\mathbf{g}}_h - \mathbf{T}_{\hat{G}} \right) \\ &= \mathbf{T}_{\hat{G}} \left(\mathbf{I} - \mathbf{W}_h \right)^2 \mathbf{T}_{\hat{G}} \end{aligned} \tag{19}$$

where $\hat{\mathbf{g}}_h = \mathbf{W}_h \mathbf{T}_{\hat{G}}$ is defined as in (17). The expected value of squared residuals given in (18) or (19) is also known as *MSE* of prediction, which in this case is

$$MSE(h) = E \left\| \mathbf{T}_{\hat{G}} - \hat{\mathbf{g}}_h \right\|^2 = E \left\| (\mathbf{I} \text{ - } \mathbf{W}_h) \mathbf{T}_{\hat{G}} \right\|^2 \tag{20}$$

It follows directly from (20) that $MSE(h)$ can be described as

$$MSE(h) = \mathbf{g}_h'(\mathbf{I}-\mathbf{W}_h)^2\mathbf{g}_h$$
$$+\sigma^2\left[n-2(\mathbf{W}_h)+\left(\mathbf{W}_h'\mathbf{W}_h\right)\right] \tag{21}$$

Hence, it follows the equation (21) that $n^{-1}RSS(h)$ is a biased estimator of the P_2 risk.

Details on the derivation of the Equation (21) can be found in the Appendix A

In practice, the equation (21) cannot be computed directly because of it depends on unknown residual variance σ_ε^2. As in linear regression, we may develop an estimator of σ_ε^2 from the residual sum of squares (18).

As a result, an estimate for σ_ε^2, as

$$\hat{\sigma}_\varepsilon^2 = \frac{RSS(h)}{n-p} = \frac{RSS(h)}{tr(\mathbf{I}-\mathbf{W}_h)^2} = \frac{RSS(h)}{DF_{RES}} \tag{22}$$

where $RSS(h)$ is defined as in (19) , and

$$\begin{aligned} DF_{RES} &= tr(\mathbf{I}-\mathbf{W}_h)^2 \\ &= n-2tr(\mathbf{W}_h)+tr\left(\mathbf{W}_h'\mathbf{W}_h\right) \end{aligned} \tag{23}$$

called the residual degrees of freedom (DF_{RES}) for pre- chosen h with any selection criteria.

As in parametric regression, DF_{RES} can be used in estimation of σ_ε^2 . Since *MSE* also has a negligible bias term, the equation (22) is an unbiased estimate of σ_ε^2 (see Ruppert et al., 2003).

As stated previously, the expected loss of a vector $\hat{\mathbf{g}}_h$ estimator can be measured by estimation of so-called R_2 risk. Our application of the results of the simulation experiments is to approximate the risk in the nonparametric regression models. Such approximates have the advantage of being simpler to optimize the practical selection of bandwidth parameters. For convenience, we will work with the scalar valued mean dispersion error.

Definition 4.1: The R_2 risk is closely related to the matrix valued mean dispersion error (MDE) of an estimator $\hat{\mathbf{g}}_h$ of $\mathbf{g}$ (see (17)). The scalar valued version of the MDE matrix is specified as

$$\begin{aligned} \text{SMDE}(\hat{\mathbf{g}}_h,\mathbf{g}) &= E(\hat{\mathbf{g}}_h-\mathbf{g})'(\hat{\mathbf{g}}_h-\mathbf{g}) \\ &= tr\left(\text{MDE}(\hat{\mathbf{g}}_h,\mathbf{g})\right) \end{aligned} \tag{24}$$

Lemma 4.1: Consider different estimators $\hat{\mathbf{g}}_h$. The mean dispersion error (MDE) of these estimators is the sum of the covariance matrix and the squared bias:

$$\begin{aligned}\mathrm{SMDE}\left(\hat{\mathbf{g}}_h\right) &= E\sum_{i=1}^{n}\left(g_i(X) - \hat{g}_{hi}(X)\right)^2 \\ &= E\left\|\mathbf{g} - \hat{\mathbf{g}}_h\right\|^2 \\ &= \left\|\left(\mathbf{I} - \mathbf{W}_h\right)\mathbf{g}\right\|^2 + \sigma_\varepsilon^2 tr\left[\mathbf{W}_h \mathbf{W}_h'\right]\end{aligned} \tag{25}$$

Proof: See Appendix B.

As shown in the lemma 4.1 the SMDE matrix decomposes into a sum of the squared bias and variance of the estimator. Hence, we can compare the quality of two estimators by looking at the ratio of their SMDE in (25). This ratio gives the following definition concerning the superiority of any two estimators.

Definition 4.2: The relative efficiency of an estimator $\hat{\mathbf{g}}_{E1}(h)$ compared to another estimator $\hat{\mathbf{g}}_{E2}(h)$ is defined by the ratio,

$$RE = \frac{R\left(\hat{\mathbf{g}}_{E1}(h), \mathbf{g}\right)}{R\left(\hat{\mathbf{g}}_{E2}(h), \mathbf{g}\right)} = \frac{\mathrm{SMDE}\left(\hat{\mathbf{g}}_{E1}(h)\right)}{\mathrm{SMDE}\left(\hat{\mathbf{g}}_{E2}(h)\right)} \tag{26}$$

where $R(.)$ denotes the scalar risk that is equivalent to the equation (24). $\hat{\mathbf{g}}_{E2}(h)$ is said to be more efficient than $\hat{\mathbf{g}}_{E1}(h)$ if $RE < 1$.

5. BANDWIDTH SELECTION CRITERIA

This section provides an overview of several criteria which have been used for smoothing parameter (or bandwidth) selection. The key idea is to select an appropriate value for the parameter h, bandwidth. As stated in previous section, the optimum h is defined as the smoothing parameter which minimizes the average of the mean square errors (AMSE), given by

$$\mathrm{AMSE}(h) = \frac{1}{n}\left\|\left(\mathbf{I} - \mathbf{W}_h\right)\mathbf{T}_{\hat{G}}\right\|^2 + \frac{\sigma_\varepsilon^2}{n} tr\left(\mathbf{W}_h^2\right)$$

where $\mathbf{W}_h$ is given in equation (16). The estimator of the error variance σ_ε^2 is defined in the equation (22).

The selection criteria are summarized as

***GCV Criterion*:** The criterion function is defined by Craven and Wahba (1979), and described as

$$\mathrm{GCV}(h) = \frac{1}{n}\frac{\left\|\left(\mathbf{I} - \mathbf{W}_h\right)\mathbf{T}_{\hat{G}}\right\|^2}{\left[n^{-1} tr\left(\mathbf{I} - \mathbf{W}_h\right)\right]^2}$$

where $\mathbf{W}_h$, as is defined in (16), is smoother matrix based on h

As in other criteria, to use GCV for parameter selection, we simply choose the parameter h giving smallest GCV over the set of parameter considered.

***AICc Criterion*:** Hurvich et. al., (1998) suggested an improved version of AIC which is called AIC$_C$, which is defined by

$$\mathrm{AIC_c}(h) = 1 + \log\left[\left\|(\mathbf{W}_h - \mathrm{I})\mathbf{T}_{\hat{G}}\right\|^2 \Big/ n\right]$$
$$+\left[2\left\{tr(\mathbf{W}_h) + 1\right\}\Big/ n - tr(\mathbf{W}_h) - 2\right]$$

***BIC Criterion*:** The BIC is also called as Schwarz Information Criterion (SIC). The criterion is expressed as

$$\mathrm{BIC}(h) = 1/n\left\|\left(\mathbf{I}\text{-}\mathbf{W}_h\right)\mathbf{T}_{\hat{G}}\right\|^2$$
$$+\left(l\,\mathrm{og}(n)/n\right)tr\left(\mathbf{W}_h\right)$$

***REML Criterion*:** The derivatives of both the REML and the GCV with respect to h can be determined quite naturally in a common form (see Reis et al., 2009). The REML score can be specified as

$$\mathrm{REML}(h) = \left\|\left(\mathrm{I}\text{-}\mathbf{W}_{\ddot{e}}\right)\mathbf{T}_{\hat{G}}\right\|^2 \Big/ n - tr\left(\mathbf{W}_h\right)$$

C$_\mathbf{P}$ Criterion: Mallows (1973) suggests the C_p criterion in the regression case. If σ^2 is recognized, an unbiased estimate of the residual sum of squares is provided by C_p criterion:

$$C_\mathrm{p}(h) = \frac{1}{n}\left\{\left\|(\mathbf{W}_h - \mathrm{I})\mathbf{T}_{\hat{G}}\right\|^2 + 2\sigma^2 tr(\mathbf{W}_h) - \sigma^2\right\}$$

Unless σ^2 is known, in practice an estimation for σ^2 can be given by (22).

RECP Criterion: A direct computation leads to the bias-variance decomposition for $R(\mathbf{g}, \hat{\mathbf{g}}_h)$:

$$R\,(\mathbf{g}, \hat{\mathbf{g}}_h) = \frac{1}{\mathrm{n}}E\left\|\mathbf{g}\,\text{-}\,\hat{\mathbf{g}}_h\right\|^2$$
$$= \frac{1}{n}\left\{\left\|(\mathbf{W}_h - \mathbf{I})\mathbf{g}\right\|^2 + \sigma^2 tr(\mathbf{W}_h\mathbf{W}_h')\right\}$$

A clear–cut explanation shows that $R(\mathbf{g}, \hat{\mathbf{g}}_h) = \left\{\mathrm{C}_p(h)\right\}$. Because the risk $R(\mathbf{g}, \hat{\mathbf{g}}_h)$ is an unknown quantity, so-called risk is now estimated by computable quantity $R\,(\hat{\mathbf{g}}_{h_p}, \hat{\mathbf{g}}_h)$. The obtained expression for $R\,(\hat{\mathbf{g}}_{h_p}, \hat{\mathbf{g}}_h)$ is

$$R\,(\hat{\mathbf{g}}_{h_\mathrm{p}}, \hat{\mathbf{g}}_h) = \frac{1}{\mathrm{n}}E\left\|\hat{\mathbf{g}}_{h_\mathrm{p}}\,\text{-}\,\hat{\mathbf{g}}_h\right\|^2$$
$$= \frac{1}{n}\left\{\left\|\left(\mathbf{W}_h - \mathbf{I}\right)\hat{\mathbf{g}}_{h_\mathrm{p}}\right\|^2 + \hat{\sigma}^2_{h_\mathrm{p}}\,tr(\mathbf{W}_h\mathbf{W}_h')\right\}$$

where $\hat{\sigma}^2_{h_\mathrm{p}}$ and $\hat{\mathbf{g}}_{h_\mathrm{p}}$ are the appropriate *pilot estimates* for σ^2 and $\mathbf{g}$, respectively. The pilot h_p selected by classical methods. (see Lee, 2003-2004)

6 SIMULATION EXPERIMENT

In this section we performed a Monte-Carlo simulation study to present and compare the kernel smoother estimators based on different bandwidth selection criteria given in section (4). To see the performance of the small, medium and large samples of each criterion we consider three censoring levels (CLs), 10%, 30%, and 50% and three samples sizes with *n*= 50, 100, and 200. The number of replication was 1000 for each of the samples. All calculations are carried out in MATLAB software. The empirical data is generated as;

$$T_i = \{g(X_i) = \sin(-4.8X)\sin(1.4X)\} + \varepsilon_i \tag{27}$$

where $X_i = \left(\dfrac{i-0.5}{n}\right)_{i=1}^{n}$ and $\varepsilon_i \sim N\left(0, \sigma^2 = 1\right)$.

6.1 Empirical Evaluations

In our simulation study 54 different configurations are carried out. Furthermore, we used the MSE values to evaluate the quality of any curve estimate ($\hat{g}_h(X_i) = (\hat{\mathbf{g}}_h)_i$):

$$MSE = \frac{1}{n}\sum_{\mathrm{i}=1}^{\mathrm{n}}\{g(X_{\mathrm{i}}) - \hat{g}_h(X_{\mathrm{i}})\}^2 \tag{28}$$

In the nonparametric regression setting, the outcomes from Monte Carlo simulation are illustrated in the following Tables and Figures.

Table 1 compares the *MSE* values connected to the nonparametric regression models with right censored data under different censoring levels and sample sizes. The main idea is that a model with a better fit denotes a minimum squared Euclidean distance between the data and fitted values, and thus it has a minimum *MSE* value.

MSE values from nonparametric regression models

Table1

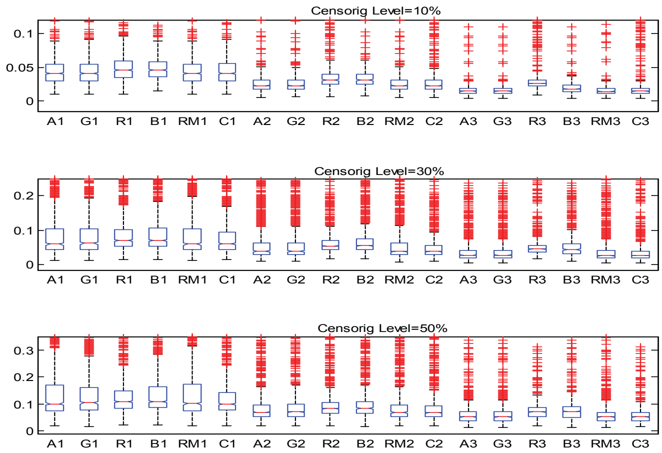
n = 50			
	CL=10%	CL=30%	CL=50%
AIC	0.067	0.129	0.201
GCV	0.066	0.126	0.187
RECP	0.063	0.111	0.170
BIC	0.070	0.135	0.206
REML	0.067	0.131	0.204
Cp	**0.057**	**0.103**	**0.165**
n = 100			
AIC	0.047	0.084	0.106
GCV	0.047	0.084	0.105
RECP	0.045	0.077	0.101
BIC	0.054	0.095	0.117
REML	0.047	0.085	0.107
Cp	**0.037**	**0.065**	**0.090**
n = 200			
AIC	0.035	0.059	0.069
GCV	0.035	0.059	0.069
RECP	0.037	0.059	0.075
BIC	0.038	0.070	0.083
REML	0.036	0.059	0.069
Cp	**0.027**	**0.045**	**0.061**

As expected, in Table 1, we obtained big MSE values for high censoring levels for all selection methods. Note also that although selection methods have good performances in general, BIC and REML methods gave bigger MSE values than AICc, GCV, RECP and Cp criteria. It means that their estimation performances are not good for bandwidth parameter under randomly right-censored data.

The effect of the censoring, as expected, tends to increase the MSE values of the estimators, losing precision as the censoring level increases. In addition, and also as expected, the MSE values are improved as the sample size increases.

Boxplots of the MSE values for estimated nonparametric models

Figure 1

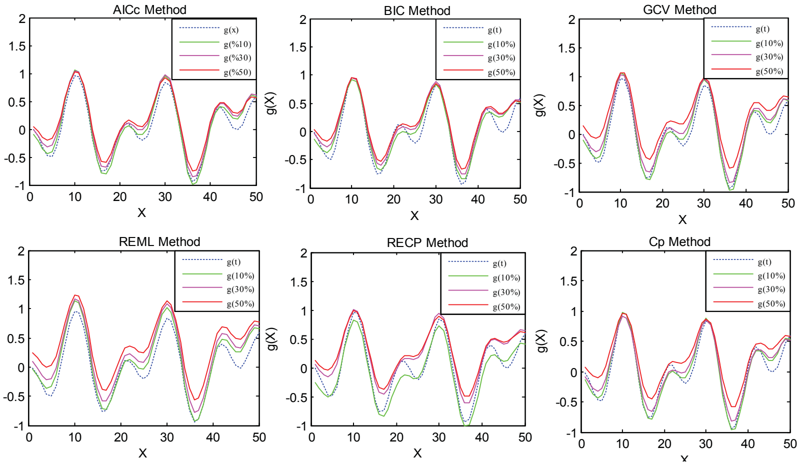


Boxplots for MSE values based on each criterion are illustrated in Figure 1. In this Figure, A1, G1, B1, R1, RM1 and C1 denote the MSE values based on AICc, GCV, BIC, RECP, REML, and Cp selection criteria for sample sizes n=50, respectively. In a similar fashion, A2, G2, R2, B2, RM2 and C2 show the MSE values depend on the same criteria but for n=100. Finally, A3, G3, R3, B3, RM3 and C3 indicate the MSE values based on the mentioned criteria but for n=200. Also, the upper panel of Figure 1 has CL=10%, medium panel CL=30%, and bottom panel CL= 50%.

As can be seen in Figure 1, as the sample size n gets large, the range of estimates are getting narrow. It can be said that the estimates from medium and large sized samples are more stable than those from small sized sample

Real observations and the true function together with its smooth curves estimated by the selection criteria under different censoring levels

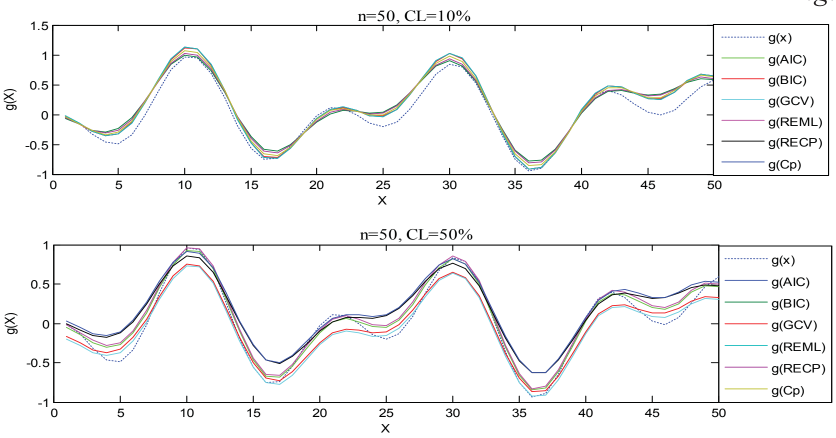
Figure 2



As can be seen from Figures 2, the estimated functions move away from the real function when censoring levels increases, regardless of the sample sizes. Also, simulation experiment results show that the quality of estimated curves is reasonable for censoring levels, CL=10% and 30%, when compared to the CL=50%.

Real data points and the true function together with its smooth curves based on six selection criteria for n=50, and 10% and 50% censoring levels, respectively.

Figure 3

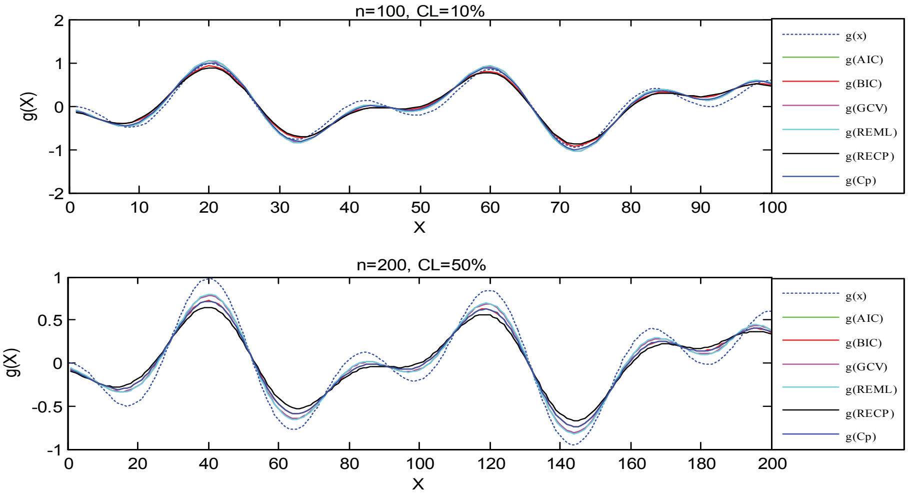


As for Figure 3, we illustrate the true function together with their curves estimated by the selection criteria for samples sizes n = 50. In this Figure, the bottom panel represents the censoring level of 10%, while upper panel shows the same curves but for censoring level of 50%. As expected, the estimated smooth curves are closer to the real function when censoring levels decreases (see upper panel of the Figure 3).

The estimated smooth curves in the Figure 4 exhibit a similar behaviour to Figure 3. That is, the curves obtained from the data with low censoring levels denote a better fit from data with high censorship. Also, the effect of censoring levels makes much more impact on the estimated curves than sample sizes.

Similar to Figure 3, but for n=100 and 200.

Figure 4



6.2 Comparing the efficiency

In this simulation study, to compare the efficiency of the selection criteria based on different censoring levels and sample sizes we obtained the relative efficiency matrix from the values of the SMDE ratios of the selection criteria. These values are computed using the equation (26) and they are given in Tables 2-3 for %10 and %50 censoring levels and all sample sizes. Outcomes correspond to %30 censoring levels are similar to the results displayed in Tables 2, they are not reported here. From Tables, we see that the relative efficiency values of Cp method are smaller than 1 for all scenarios. Hence, it can be said that Cp is more efficient than the other selection criteria for all sample sizes and censoring levels.

Efficiency values of selection criteria for 10% censoring level and all sample sizes

Table 2

	AIC	**GCV**	**RECP**	**BIC**	**REML**	**Cp**
			n = 50, CL=10%			
AIC	**1.000**	0.998	0.907	1.026	1.027	0.814
GCV	1.001	**1.000**	0.909	1.028	1.029	0.815
RECP	1.101	1.099	**1.000**	1.130	1.132	0.897
BIC	0.974	0.972	0.884	**1.000**	1.001	0.793
REML	0.973	0.971	0.883	0.998	**1.000**	0.792
Cp	1.227	1.222	1.114	1.260	1.262	**1.000**
			n=100, CL=10%			
AIC	**1.000**	1.000	0.930	1.118	1.026	0.772
GCV	1.000	**1.000**	0.930	1.118	1.026	0.772
RECP	1.075	1.075	**1.000**	1.202	1.103	0.830
BIC	0.894	0.894	0.832	**1.000**	0.918	0.690
REML	0.974	0.974	0.906	1.089	**1.000**	0.752
Cp	1.295	1.295	1.204	1.447	1.329	**1.000**
			n=200, CL=10%			
AIC	**1.000**	1.000	0.991	1.056	1.024	0.742
GCV	1.000	**1.000**	0.991	1.056	1.024	0.742
RECP	1.008	1.008	**1.000**	1.065	1.032	0.748
BIC	0.946	0.946	0.938	**1.000**	0.969	0.702
REML	0.976	0.976	0.968	1.031	**1.000**	0.724
Cp	1.347	1.347	1.336	1.423	1.380	**1.000**

Findings of the simulation study may be summarized as follows:

- It is observed that the estimator using the Cp choice of bandwidth parameter h dominates the other estimators for all scenarios.
- Inspection of the relative efficiency values also reveal that for %50 censoring rate AIC criterion converges 1 highest at rates when sample size is large, n =200.
- Notice that for all samples sizes and CL=%10, the AIC and GCV produce the same relative efficiency values, whereas RECP gives the similar values to these criteria.
- Simulated relative efficiencies of BIC REML are not good and decreases dramatically with sample sizes, especially for %10 censoring levels.

Efficiency values of selection methods for %50 censoring level and all sample sizes

Table 3

	AIC	GCV	RECP	BIC	REML	Cp
n = 50, CL=50%						
AIC	1.000	0.953	0.848	1.024	1.017	0.822
GCV	1.048	1.000	0.889	1.074	1.066	0.862
RECP	1.179	1.124	1.000	1.207	1.199	0.969
BIC	0.976	0.931	0.828	1.000	0.993	0.803
REML	0.983	0.937	0.833	1.006	1.000	0.808
Cp	1.216	1.159	1.031	1.245	1.236	1.000
n=100, CL=50%						
AIC	1.000	0.998	0.932	1.105	1.014	0.830
GCV	1.001	1.000	0.934	1.107	1.016	0.832
RECP	1.071	1.070	1.000	1.184	1.087	0.890
BIC	0.905	0.903	0.844	1.000	0.918	0.751
REML	0.985	0.983	0.919	1.089	1.000	0.818
Cp	1.203	1.201	1.123	1.330	1.221	1.000
n=200, CL=50%						
AIC	1.000	1.000	1.055	1.184	1.006	0.853
GCV	0.999	1.000	1.055	1.184	1.006	0.853
RECP	0.947	0.947	1.000	1.122	0.953	0.808
BIC	0.844	0.844	0.890	1.000	0.849	0.720
REML	0.993	0.993	1.048	1.177	1.000	0.848
Cp	1.171	1.171	1.236	1.388	1.179	1.000

In the next section, we used a censored real data to see the process of the selection criteria.

7. REAL DATA EXAMPLE

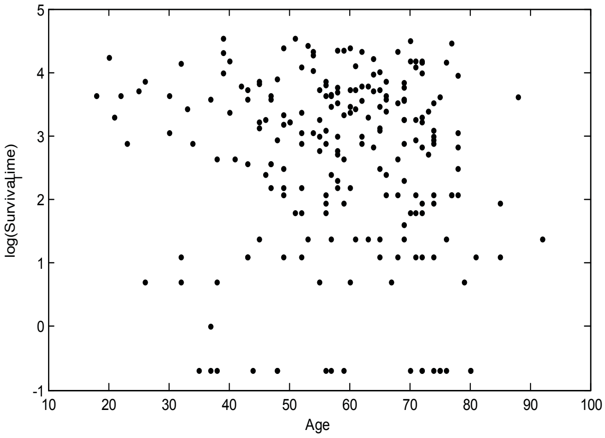
To motivate the problem of the kernel type estimation procedure in nonparametric regression model with censored data, we used bowel cancer data obtained from cancer patients in Izmir city of Turkey. In here the logarithm of the survival times is considered as response (logT), while patient's age is used as covariate (*X*).

As seen from inspection of Figure 5, there is no strong evidence of a linear relationship between survival times and age. To see the relationship between survival time and age, the residuals are plotted against age in Figure 6. The nonlinearity is now more evident, especially because a scatterplot smooth has been added. This suggests that a nonparametric regression approach will be beneficial.

.

Scatterplot of age and lifetime data

Figure 5



The mentioned nonparametric regression model can be expressed as follows:

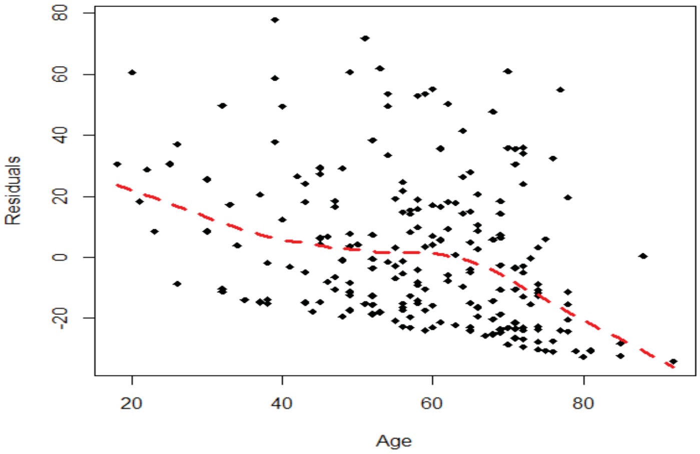
$$\log\left(survival\,times\right)_i = g(age)_i + \varepsilon_i,\ i = 1,..,218$$

where $survival\,times$ and age are defined as above, response and covariate, respectively.

As previously mentioned, the key idea is to estimate the unknown function $g(age)$. Various kernel estimates of these functions are obtained by using six selection criteria choice of bandwidth parameter, and showed in Figure 7.

Scatterplot residuals from regression of survival times on age

Figure 6



As shown in Figure 7, there are six smooth functions that show the general trend of the data**.** These regression functions or smooth curves are also the plot of

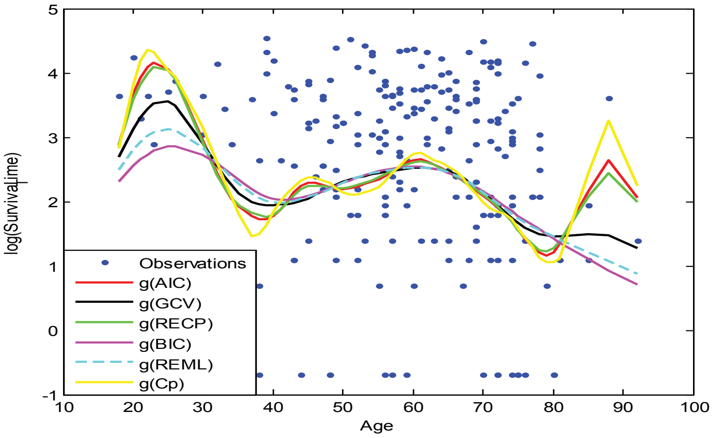
$$\hat{\mathbf{g}}_h = \left(\hat{g}_h(age_1),...,\hat{g}_h(age_{218})\right)'$$

using (17), different nonparametric estimates of the effect of age variable on $survival\,times$.

It is displayed six different smoothed curves for the kernel type estimators using the AICc, GCV, RECP, BIC, REML and Cp choice of bandwidth parameter h, respectively. The MSE values obtained from these kernel smoothing fits are 2.223, 2.225, 2.219, 2.233, 2.225 and 2.217, respectively. Thus, kernel fits of the nonparametric model obtained by AICc, GCV, RECP, BIC, and REML give similar performance, while Cp denotes a good performance in the estimation procedure.

Real observations and their smoothed curves obtained by six different kernel type estimators using six criteria choice of bandwidth parameter

Figure 7



8. CONCLUDING REMARKS

In this paper, we discussed the estimating the nonparametric regression function using kernel smoothing when the responses are subject to randomly right censoring data. Most important problem connected with the use of a kernel estimator is the selection of a good value of bandwidth parameter. In order to select this parameter, it is considered most widely used six different bandwidth selection criteria. Note also that we have focused on estimating the bandwidth that minimizes the P_2 risk or MSE. Thus we obtained six different kernel estimators by using bandwidth parameters that minimizes the selection criteria.

This study is mainly conducted to evaluate the performances of the selection methods mentioned above. For these purposes, we used both simulated and real survival data examples.

Consequently, as expected, we obtained big MSE values for high censoring levels for all selection methods. Also, as expected, the estimated smooth curves are closer to the real function when censoring levels decreases. As for selection criteria, it is observed that Cp has had the best empirical performance. However, BIC has produced the worst result.

Finally, by considering the real data and simulation findings given in the above, the following suggestions have to be taken into account:

- Cp criterion is recommended as being the best selection criteria for all sample sizes and all censoring levels.
- For especially small sample sizes, the use RECP and GCV would be more appropriate.
- For large samples, we propose the implementation of Cp or AICc criteria.

Appendix A

We begin by considering the general definition of quadratic form, Theorem 1 and Lemmas 1-2 for proof of the equations (21)

Definition A1: Let $\mathbf{W}_h = \left[w_{ij} \right]$ be a positive semi-definite and symmetrical $n \times n$ matrix depend on the h and $\boldsymbol{\varepsilon} = \left(\varepsilon_1, ..., \varepsilon_n \right)'$ be $n \times 1$ a vector of random variables. Then

$$\text{q} = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \varepsilon_i \varepsilon_j = \boldsymbol{\varepsilon}' \mathbf{W}_h \boldsymbol{\varepsilon} \tag{A1}$$

is a called a quadratic form in $\boldsymbol{\varepsilon}$ and $\mathbf{W}_h$ is a called the matrix of a quadratic form.

Theorem A1: If $E\left(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}\right) = Cov\left(\boldsymbol{\varepsilon}\right) = \boldsymbol{\Sigma} = \left(\sigma_{ij}\right)$, and $E\left(\boldsymbol{\varepsilon}\right) = 0$, then

$$E\left(\boldsymbol{\varepsilon}' \mathbf{W}_h \boldsymbol{\varepsilon}\right) = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}\, \sigma_{ij} = tr\left(\mathbf{W}_h \boldsymbol{\Sigma}\right)$$

where $tr(.)$ denotes trace of the matrix (.)

Proof

$$E\left[\left(\boldsymbol{\varepsilon}-E(\boldsymbol{\varepsilon})\right)'\mathbf{W}_h\left(\boldsymbol{\varepsilon}-E(\boldsymbol{\varepsilon})\right)\right]$$

$$=E\left[\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}\left(\varepsilon_i-E(\varepsilon_j)\right)'\left(\varepsilon_i-E(\varepsilon_j)\right)\right]$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}E\left[\left(\varepsilon_i-E(\varepsilon_j)\right)'\left(\varepsilon_i-E(\varepsilon_j)\right)\right]$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}Cov\left(\varepsilon_i,\varepsilon_j\right)=tr\left(\mathbf{W}_h\Sigma\right)$$

as claimed.

Theorem A2: Let $\boldsymbol{\varepsilon}$ be a $n\times 1$ random vector with $E(\boldsymbol{\varepsilon})=\boldsymbol{\mu}$ and $Cov(\boldsymbol{\varepsilon})=\Sigma=\left(\sigma_{ij}\right)$. Let $\mathbf{W}_h$ be a $n\times n$ constant matrix. Then, the expected value of the equation (A1)

$$E\left(\boldsymbol{\varepsilon}'\mathbf{W}_h\boldsymbol{\varepsilon}\right)=\boldsymbol{\mu}'\mathbf{W}_h\,\boldsymbol{\mu}+tr\left(\mathbf{W}_h\Sigma\right) \qquad \text{(A2)}$$

Proof

It is well known that for $i\neq j$,

$\sigma_{ij}=E\left(\varepsilon_i\varepsilon_j\right)-\mu_i\mu_j$

and that for $i=j$,

$\sigma_{ij}=\sigma_{ii}=E\left(\varepsilon_i^2\right)-\mu_i^2=\sigma_i^2$

According to (A1), the expected value of the quadratic form $\boldsymbol{\varepsilon}'\mathbf{W}_h\boldsymbol{\varepsilon}$ in expanded form as

$$\mathrm{E}(\mathrm{q})=E\left(\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}\varepsilon_i\varepsilon_j\right)=\sum_{i=1}^{n}\sum_{j=1}^{n}E\left(w_{ij}\varepsilon_i\varepsilon_j\right)$$

Since $\sigma_{ij}=E\left(\varepsilon_i\varepsilon_j\right)-\mu_i\mu_j$, we obtain

$$E\left(\varepsilon_i\varepsilon_j\right)=\sigma_{ij}+\mu_i\mu_j$$

Substituting,

$$\mathrm{E}(\mathrm{q})=\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}E\left(\varepsilon_i\varepsilon_j\right)=\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}\left(\sigma_{ij}+\mu_i\mu_j\right)$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}\,\sigma_{ij}+\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}\,\mu_i\mu_j \qquad \text{(A3)}$$

Note also that the terms σ_{ij} are the elements of the variance-covarinace matrix Σ. This matrix is a symmetric matrix whose ith element is the variance of ε_i and whose (*ij*)th off-diagonal element is the covariance between ε_i and ε_j .

It follows from (A1), and theorem 1 that the equation (A3) is equivalent to

$$E\left(\boldsymbol{\varepsilon}'\mathbf{W}_h\boldsymbol{\varepsilon}\right)=tr\left(\mathbf{W}_h\,\Sigma\right)+\boldsymbol{\mu}'\mathbf{W}_h\,\boldsymbol{\mu}$$

This completes the proof of the theorem 2.

Again, let's consider the equation (18)

$$RSS(h)=\left(\hat{\mathbf{g}}_h-\mathbf{T}_{\hat{G}}\right)'\left(\hat{\mathbf{g}}_h-\mathbf{T}_{\hat{G}}\right)$$
$$=\mathbf{T}_{\hat{G}}\left(\mathbf{I}-\mathbf{W}_h\right)^2\mathbf{T}_{\hat{G}}$$

Thus, from Theorems A1-A2 connected with quadratic form, the expected value of the $RSS(h)$ is stated as

$$\mathrm{E}\left[RSS(h)\right]=MSE(h)=E\left\|\hat{\mathbf{g}}_h-\mathbf{T}_{\hat{G}}\right\|^2$$
$$=E\left\|\left(\mathbf{I}-\mathbf{W}\right)\mathbf{T}_{\hat{G}}\right\|^2$$
$$=E\left\|\mathbf{T}_{\hat{G}}\left(\mathbf{I}-\mathbf{W}\right)\left(\mathbf{I}-\mathbf{W}\right)\mathbf{T}_{\hat{G}}\right\|$$
$$=\mathbf{g}_h'\left(\mathbf{I}-\mathbf{W}_h\right)^2\mathbf{g}_h+\sigma^2\left(tr\left(\mathbf{I}-\mathbf{W}_h\right)^2\right)$$
$$=\mathbf{g}_h'\left(\mathbf{I}-\mathbf{W}_h\right)^2\mathbf{g}_h+n\sigma^2-2\sigma^2\left(\mathbf{W}_h\right)+\sigma^2\left(\mathbf{W}_h{}'\mathbf{W}_h\right)$$
$$=\mathbf{g}_h'\left(\mathbf{I}-\mathbf{W}_h\right)^2\mathbf{g}_h+\sigma^2\left[n-2\left(\mathbf{W}_h\right)+\left(\mathbf{W}_h{}'\mathbf{W}_h\right)\right]$$

as defined in the equation (21).

Appendix B

Proof of the Lemma 4.1

$$\mathrm{SMDE}=E\left\|\mathbf{g}-\hat{\mathbf{g}}_h\right\|,\ \text{where}\ \hat{\mathbf{g}}_h=\mathbf{W}_h\mathbf{T}_{\hat{G}}$$

Then the scalar valued version of the MDE matrix can be specified as

$$\mathrm{SMDE}\left(\hat{\mathbf{g}}_h\right)=\sum_{i=1}^{n}\left[\left(g_i(X)-E\left(\hat{g}_{hi}(X)\right)\right)\right]^2$$

$$+Cov\left[\hat{g}_{hi}(X)\right]=\sum_{i=1}^{n}\left[g_i(X)-E\left(\mathbf{W}_h\mathbf{T}_{\hat{G}}\right)_i\right]^2$$

$$+Cov\left[\left(\mathbf{W}_h\mathbf{T}_{\hat{G}}\right)_i\right]=\sum_{i=1}^{n}\left[g_i(X)-E\left(\mathbf{W}_h\mathbf{T}_{\hat{G}}\right)_i\right]^2$$

$$+Cov\left[\left(\mathbf{W}_h\mathbf{T}_{\hat{G}}\right)\right]_{ii}$$

$$=\left\|\left(\mathbf{I}-\mathbf{W}_h\right)\mathbf{g}\right\|^2+tr\left[Cov\left(\mathbf{W}_h\mathbf{T}_{\hat{G}}\right)\right]$$

$$=\left\|\left(\mathbf{I}-\mathbf{W}_h\right)\mathbf{g}\right\|^2+tr\left[\mathbf{W}_hCov\left(\mathbf{T}_{\hat{G}}\right)\mathbf{W}'_h\right]$$

Assume that $Cov\left(\mathbf{T}_{\hat{G}}\right)=\sigma_\varepsilon^2\mathbf{I}_n$ yields

$$\mathrm{SMDE}\left(\hat{\mathbf{g}}_h\right)=\left\|\left(\mathbf{I}-\mathbf{W}_h\right)\mathbf{g}\right\|^2+\sigma_\varepsilon^2 tr\left[\mathbf{W}_h\mathbf{W}'_h\right]$$

REFERENCES*:*

1. **Craven., P.,** and **Wahba, G.**,1979, *Smoothing noisy data with spline functions.*, Num. Math., 31,377-403.
2. **Eubank, R. L.**, 1988, *Spline Smoothing and Nonparametric Regression*. Macal Dekker, New York,
3. **Hardle, W., Müller, M.**, 1997, *Multivariate and Semiparametric Kernel Regression*, University of Humboldt Institute of statistics and Econometrics Berlin, Germany.
4. **Hardle, W.**, 1990, *Applied Nonparametric Regression*, Cambridge University Press.
5. **Hurvich, C., M., Simonoff, J. S.**, and **Tasi, C. L**, 1998, *Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion.*, J.R. Statist. Soc. B., 60271-293
6. **Koul, H., Susarla, V., Van Ryzin, J.,** 1981, *Regression Analysis With Randomly Right-Censored Data*, The Annals Of Statistics, 1276-1285.
7. **Kaplan E. L., Meier, P.**, 1958, *Nonparametric Estimation From İncomplete Observations*, Journal Of The American Statistical Association, Vol. 53(282), 457-481.
8. **Leurgans, S.**, 1987, *Linear Models*, Random Censoring and Synthetic Data, Biometrika Vol. 74, 301-309.
9. **Mallows', C.**, 1973, *Some comments on Cp*, Technometrics, 15,661-675.
10. **Nadaraya, E. A.**, 1964, *On estimating regression*. Theory of Probability and its Applications, pp.141–142.
11. **Peterson, A. V. Jr.**, 1977, *Expressing the Kaplan-Meier estimator as a function of empirical subsurvival functions*. J. Amer. Statist. Assoc. 72, 854-858.
12. **Reis, P. T., & Ogden, R. T.**, 2009, *Smoothing parameter selection for a class of semiparametric linear models*, J. R. Stat. Soc. B 71 (2009), pp. 505–523.
13. **Ruppert D, Wand M, Carroll R**, 2003, *Semiparametric Regression*. Cambridge University Press, Cambridge, UK.
14. **Schwarz, G.**, 1978, *Estimating the dimension of a model*, The Annals of Statistics, 6, 461-464.
15. **Stute, W., Wang, J.-L.**, 1993, *The Strong Law Under Random Censorship*, Annals of Statistics, 21,146-156.

16. **Stute, W.**, 1993, *Consistent Estimation Under Random Censorship When Covariables Are Present*, Journal Of Multivariate Analysis,45,89-103.
17. **Terrell, G.R., Scott, D.W.,** 1992, *Variable kernel density estimation*, Annals of Statistics, 20, 1236-1265.
18. **Wang, Q.-H., Li, G.**, 2002, *Empirical Likelihood Semiparametric Regression Analysis Under Random Censorship*, Journal Of Multivariate Analysis, Vol. 83(2), 469-486.
19. **Watson, G. S.**,1964, *Smooth regression analysis*, Sankhya, Series A, vol.26, 359-72.
20. **Wong, W. H.**, 1983, *On the consistency of cross-validation in kernel nonparametric regression*, Annals of Statistics, Vol.11(4),1136-1141.
21. **Van der Vaart, A. W.**, 1998, *Asymptotic Statistics*, Cambridge University press.
22. **Vieu, P.**, 1991, *Nonparametric regression: optimal local bandwidth choice*, Journal of he Royal Statistical Society Vol. 53(2), 453-464.
23. **Zheng, Z. K.,** 1984, *Regression Analysis wtih Censored Data*, Ph.D Dissertation, Univ of Colombia
24. **Lee, Thomas C. M.**, 2004, *Improved smoothing spline regression by combining estimates of different smoothness*, Statistics & Probability Letters, 67, 133-140.
25. **Lee, Thomas C. M.**, 2003, *Smoothing parameter selection for smoothing splines: a simulation study*, Computational Statistics & Data Analysis, 42, 139-148.