Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Smoothing parameter selection in semiparametric regression models with censored data

Dursun Aydin *, Ersin Yilmaz

Department of Statistics, Mugla Sitki Kocman University, Mugla, Turkey

ARTICLE INFO

Article history:
Received 2 May 2017
Received in revised form
21 July 2017
Accepted 26 July 2017

Keywords:
Right-censored data
Semi-parametric regression
Penalized spline
Kaplan-meier estimation

ABSTRACT

In this paper, we introduce penalized spline estimators for the unknown function and a parameter vector in a semiparametric regression model with right censored data. In order to obtain this estimator accurately and efficiently, we used penalized spline method based on three important selection criteria such as corrected Akaike's information criterion (AIC), generalized cross-validation (GCV) and Mallows' Cp criterion (MCp). The purpose of the study is to illustrate the performance of penalized regression spline method for estimating the right-censored data and also comparing the mentioned three selection methods in selection of smoothing parameter. The ideas that expressed in this study are demonstrated in a real cancer patients' data and a Monte Carlo simulation using different censoring levels and sample sizes. Thus, the appropriate selection criteria are provided for a suitable smoothing parameter selection. Cp gave satisfying results for this study.

1. Introduction

Consider the semiparametric regression model (Eq. 1);

$$T_i = X_i\beta_i + f(Z_i) + \varepsilon_i, i = 1, \dots, n \quad (1)$$

where T_i's are response observations, $X_i = (X_{i1}, \dots, X_{ip})$is p-dimensional vector, Z_i's are observations of an extra univariate variable, $\beta = (\beta_1, \dots, \beta_p)'$is an unknown p-dimensional parameter vector to be estimated, $f(.)$is an unknown smooth function and ε_i's are random error terms with zero mean and variance σ^2. There are many approaches to estimate β and f in the model (1) with uncensored data. One of the primary approaches is the penalized splines method discussed by Eilers and Marx (2010) and Ruppert et al. (2003). Furthermore, Hall and Opsomer (2005), and Liang (2006) gave some theoretical results on penalized splines.

In our study, we are interested in estimating the parameter vector β and unknown function f in model (1), when the T_i's are observed incompletely and right censored by a random variable C_i , $i = 1, \dots, n$ while $X_i$and $Z_i$are observed completely. Therefore, incomplete observations $\{X_i, Z_i, T_i\}$ are transformed to complete observations $\{X_i, Z_i, L_i, \delta_i, 1 \le i \le n\}$ as follows

$$L_i = min(T_i, C_i)$$

and

$$\delta_i = I(T_i \le C_i), \ i = 1, \dots, n \quad (2)$$

where L_i 's are the adjusted response observations with unknown distribution M, C_i's are the values of the censoring variable and δ_i's are the values of the censoring indicator. As can be seen from the Eq. 2, censoring indicator provides the censoring information; if ith observation is complete then δ_i takes 1 otherwise 0. Also, we assume that T and C are independent random variables with unknown distributions F and G, respectively.

In this paper, we focus on the model (1) when the response variable is incompletely observed (or censored data). In the literature, for the censored linear and nonlinear regression models, appropriate estimators are defined by replacing incomplete observations with synthetic data (Buckley and James, 1979; Koul et al., 1981; Lai and Ying, 1992) for simplicity, we consider the transformed versions of the censored observations, called as synthetic data, proposed by Koul et al. (1981). We apply a penalized regression spline estimator to the synthetic data to determine a proper smoothing parameter selection criterion. The above estimator is

* Corresponding Author.
Email Address: duaydin@hotmail.com (D. Aydin)
https://doi.org/10.21833/ijaas.2017.08.024

a generalization of the well-known penalized spline estimator for the model (1).

In our study, the main difference is that we consider a randomly right censored semiparametric regression model, estimated by using several smoothing parameter selection criteria. The basic idea is to find a useful selection criterion that provides a good approximation to the model (1). Also it is aimed that the comparison between the performances of these different selection criteria.

In case of the censored data, a number of authors have studied a semiparametric regression model (1). Examples of this study include Chen and Khan (2001), Qin and Jing (2000), Wang and Li (2002), Orbe et al. (2003), Lu and Cheng (2007), and Zhou and Liang (2009). Also note that, estimation procedure for right-censored response observations is proposed by Kaplan and Meier (1958) and then Miller (1976) proposed Kaplan-Meier weights for linear regression model estimation using with K-M estimator. Also, Stute (1999) studied on nonlinear censored regression model with K-M weights and inspected its theoretical and asymptotical properties. In addition to these, Koul et al. (1981) suggested synthetic data transformation with using K-M estimator as an alternative for K-M weights.

The rest of the paper is designed as follows. Section 2 gives information about the preliminaries and methodology. The ideas on the features of the model and the proposed estimator are given in section 3. In Section 4, it is discussed the smoothing parameter selection criteria. Section 4 provides a real data example and a Monte Carlo simulation study. Finally, the conclusions are presented in the Section 5.

2. Preliminaries and methodology

Let F, G and M be the cumulative distribution functions of Y, C and L respectively. These are assumed to be positive random variables with distribution functions

$$F(s|X,Z) = P(T \le s|X,Z), G(s|X,Z) = P(C \le s|X,Z)$$

and

$$M(s|X,Z) = P(L \le s|X,Z)\ (s \in R) \quad (3)$$

from Eq. 3 their unknown survival functions are

$$\bar{F}(s|X,Z) = P(T > s|X,Z), \bar{G}(s|X,Z) = P(C > s|X,Z)$$

and

$$\bar{M}(s|X,Z) = P(L > s|X,Z)\ (s \in R) \quad (4)$$

Under assumption that T and C are independent, survival function M can be written as

$$\bar{M}(s|X,Z) = \bar{F}(s|X,Z).\bar{G}(s|X,Z)P(T > s|X,Z).P(C > s|X,Z) = P(L > s|X,Z)$$

To assess the statistical accuracy of the model which is computed from censored data set, we also assume that

$$\tau_F = \inf[s: \bar{F}(s|X,Z) = 1], \tau_G = [s: \bar{G}(s|X,Z) = 1]$$
$$\tau_M = \inf[s: \bar{M}(s|X,Z) = 1] \quad (5)$$

where $\tau_F<\infty$, F and G distributions have no jumps and G is continuous. It follows from (3) that, there is a probability distribution for T at each possible values for X and Z. From Eq. 5, for given s point, the mean of this distribution can be defined as

$$\int_0^\infty F(s|X,Z)ds = \int_0^{\tau_F} F(s|X,Z)ds = E(T|X = x, Z = z) \quad (6)$$

Because of the nature of the censored data, the estimation of semiparametric model cannot be directly computed by the traditional methods here. To get rid of this problem Koul et al. (1981) proposed a synthetic data transformation. Normally, right-censored response variable (Y) and updated response variable (T) have different expectations because of censoring but synthetic data transformation provide a regularization for solving this problem (see Appendix A). That is, the censored response observations L_i's are converted to a synthetic response values T_{iG} according to the rule

$$T_{iG} = \frac{\delta_i L_i}{\bar{G}(L_i)} = \frac{\delta_i L_i}{1-G(L_i)} \quad (7)$$

where $G(.)$ is the common distribution of the censoring variable C_i as mentioned in the introduction to this section. In the censored data applications, however, the censoring distribution G is usually unknown. For this reason, to estimate the components of the model (1) the ordinary methods cannot be used directly here. To overcome this problem Koul et al. (1981) proposed to replace G with its Kaplan Meier estimator, given by

$$\hat{\bar{G}}(s) = 1 - \hat{G}(s) = \prod_{i=1}^{n}\left(\frac{n-i}{n-i+1}\right)^{I(L_i \le s, \delta_i = 0)}, s \ge 0 \quad (8)$$

where $L'_{(i)}s, L_{(1)} \le L_{(2)} \le \cdots \le L_{(2)}$ are the ordered values of variable $L_{(i)}$ and $\delta_{(i)}$values are the ordered associated with values of $L_{(i)}$. Substituting $\hat{\bar{G}}(s) = 1 - \hat{G}(s)$ for $G(.)$ in (7), we obtain the following synthetic data:

$$T_{iG} = \frac{\delta_i L_i}{\hat{\bar{G}}(L_i)} = \frac{\delta_i L_i}{1-\hat{G}(L_i)} \quad (9)$$

We will now see how to estimate the smooth function $f(Z)$ and the regression coefficients vector β, based on the above synthetic observations. For these purposes, several estimation methods, such as smoothing spline, kernel smoothing and regression spline, are improved in the literature. For convenience, we use regression spline method to estimate the unknown regression function and parametric coefficients in the model (1). This estimate procedure is explained in the following section.

2.1. Derivation of the estimator

Now we consider the ideas described above to apply the penalized spline method to the case of censored data. In this method, $f(Z)$ is approximated by rth degree penalized spline with truncated polynomial basis (Ruppert et al., 2003):

$$f(Z;\alpha) = \alpha_0 + \alpha_1 Z + \alpha_2 Z^2 + \cdots + \alpha_r Z^r + \sum_{k=1}^{K} b_k (Z - \kappa_k)_+^r \quad (10)$$

where $r \geq 1$is an integer that indicates the degree of penalized spline and $(Z-\kappa_k)_+ = (Z-\kappa_k)$ if $(Z > \kappa_k)$ and 0 otherwise. Also $\kappa_1 < \cdots < \kappa_K$ are the specifically selected knots $\{\min(Z_i) \leq \kappa_1 < \cdots < \kappa_K \leq \max(Z_i)\}$.

There are several studies about selection of number of knots. In this study we selected knots according to full search algorithm (FSA) method which is expressed in Ruppert et al. (2003).

Using the above truncated polynomial and (9), it follows that the censored regression model (1) can be modified as a mixed model, given by

$$T_{i\hat{G}} = \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \alpha_0 + \alpha_1 Z_i + \cdots + \alpha_r Z_i^r + \sum_{k=1}^{K} b_k (Z_i - \kappa_k)_+^r + \varepsilon_{i\hat{G}} \quad (11)$$

where $i = 1, \ldots, n$ and $T_{i\hat{G}}$'s are the synthetic response observations $\varepsilon_{i\hat{G}}$'s are random error terms for given G and $n \to \infty$, $E(\varepsilon_{i\hat{G}}) \cong 0$ (see Appendix B). Thus, we can fit model (1) using penalized spline through the mixed model (11). In a matrix and vector form, the model (11) can be rewritten as Brumback et al. (1999) and Ruppert et al. (2003)

$$\mathrm{T}_{\hat{G}} = \mathrm{XB} + \mathrm{Ab} + \varepsilon_{\hat{G}} \quad (12)$$

Where

$$\mathrm{B} = (\beta_1, \ldots, \beta_p, \alpha_0, \ldots, \alpha_r)', \ \mathrm{b} = (b_1, \ldots, b_K)', \mathrm{T}_{\hat{G}} = (t_{1\hat{G}}, \ldots, t_{n\hat{G}})$$

and

$$\varepsilon_{\hat{G}} = (\varepsilon_{1\hat{G}}, \ldots, \varepsilon_{n\hat{G}}).$$

X and A are design matrices, given by

$$\mathbf{X} = \begin{bmatrix} X_{1i} & \cdots & X_{pi} & 1 & \cdots & Z_1^r \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ X_{1n} & \cdots & X_{pn} & 1 & \cdots & Z_n^r \end{bmatrix}, \mathbf{A} = \begin{bmatrix} (Z_1-\kappa_1)_+^r & \cdots & (Z_1-\kappa_K)_+^r \\ \vdots & & \vdots \\ (Z_n-\kappa_1)_+^r & \cdots & (Z_1-\kappa_K)_+^r \end{bmatrix}$$

The penalized spline estimators $\hat{\mathrm{B}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\alpha}_1, \ldots, \hat{\alpha}_r)', \hat{\mathrm{b}} = (\hat{b}_1, \ldots, \hat{b}_K)'$of (B, b) are obtained by minimizing the penalized least-squares criterion

$$PRSS(\mathrm{B};\mathrm{b}) = (\mathrm{T}_{\hat{G}} - \mathrm{XB} - \mathrm{Ab})'(\mathrm{T}_{\hat{G}} - \mathrm{XB} - \mathrm{Ab}) + \lambda \mathrm{b}'\mathrm{Db} \quad (13)$$

where $\mathrm{D} = diag(0_{p+1}, 1_K)$ that is D is a diagonal penalty matrix whose first (p + 1) elements are 0, and the remaining elements are 1.

The part λb′Db in (13) is called a penalty term because it penalizes curvatures in the function f, thus yielding a smoother result. The amount of penalty is controlled by a smoothing parameter$\lambda > 0$. In general, large values of λ produce smoother estimators while smaller values produce more wiggly estimators. As can be seen from here, the parameter λ plays a key role in estimating of the model (11). Also, one of our tasks is to select an optimal value of the λ in here. This problem is discussed in section (4).

Minimization of the criterion (13) leads to the system of equations

$$\begin{bmatrix} \mathrm{X}'\mathrm{X} & \mathrm{X}'\mathrm{A} \\ \mathrm{A}'\mathrm{X} & (\mathrm{A}'\mathrm{A} + \lambda \mathrm{D}) \end{bmatrix} \begin{bmatrix} \mathrm{B} \\ \mathrm{b} \end{bmatrix} = \begin{bmatrix} \mathrm{X}' \\ \mathrm{A}' \end{bmatrix} \mathrm{T}_{\hat{G}} \quad (14)$$

Using the Eq. 14, we obtain the regression spline estimator for the coefficients vector **b** in the model (12), given by

$$\hat{\mathrm{b}} = (\mathrm{A}'\mathrm{A} + \lambda \mathrm{D})^{-1}\mathrm{A}'(\mathrm{T}_{\hat{G}} - \mathrm{XB}) \quad (14a)$$

Then, substituting this estimator of b into (13), we obtain

$$\mathrm{B} = (\mathrm{X}'\mathrm{U}^{-1}\mathrm{X})^{-1}\mathrm{X}'\mathrm{U}^{-1}\mathrm{T}_{\hat{G}} \quad (14b)$$

where $\mathrm{U}^{-1} = \mathrm{I} - \mathrm{A}(\mathrm{A}'\mathrm{A} + \lambda \mathrm{D})^{-1}\mathrm{A}$ and thus, the vector of fitted values is given by

$$\hat{\mathrm{T}}_{\hat{G}} = (\mathrm{X}\hat{\mathrm{B}} + \mathrm{A}\hat{\mathrm{b}}) = \mathrm{H}_\lambda \mathrm{T}_{\hat{G}}$$

where

$$\mathrm{H}_\lambda = \mathrm{A}(\mathrm{A}'\mathrm{A} + \lambda \mathrm{D})^{-1}\mathrm{A}' + (\mathrm{I} - \mathrm{A}(\mathrm{A}'\mathrm{A} + \lambda \mathrm{D})^{-1}\mathrm{A}')\mathrm{X}(\mathrm{X}'\mathrm{U}^{-1}\mathrm{X})^{-1}\mathrm{X}'\mathrm{U}^{-1} \quad (14c)$$

2.2. Estimation of variance

In practice smoothing parameter λ depends on variance of the error terms $\varepsilon_{\hat{G}} = \mathrm{T}_{\hat{G}} - \mathrm{A}\hat{\mathrm{b}} - \mathrm{X}\hat{\mathrm{B}}$. As in linear regression, we can develop an estimator for σ^2 from the error or residuals sum of squares (RSS)

$$RSS = (\mathrm{T}_{\hat{G}} - \hat{\mathrm{T}}_{\hat{G}})'(\mathrm{T}_{\hat{G}} - \hat{\mathrm{T}}_{\hat{G}}) = (\mathrm{T}_{\hat{G}} - \mathrm{H}_\lambda \mathrm{T}_{\hat{G}})'(\mathrm{T}_{\hat{G}} - \mathrm{H}_\lambda \mathrm{T}_{\hat{G}}) = \|(\mathrm{I} - \mathrm{H}_\lambda)\mathrm{T}_{\hat{G}}\|^2 \quad (15)$$

where $\mathrm{H}_\lambda$is a smoother matrix, as in defined in(14c). Thus, an estimator of σ^2 is obtained as

$$\hat{\sigma}^2 = \frac{RSS}{tr(\mathrm{I}-\mathrm{H}_\lambda)^2} = \frac{\|(\mathrm{I}-\mathrm{H}_\lambda)\mathrm{T}_{\hat{G}}\|^2}{tr((\mathrm{I}-\mathrm{H}_\lambda)'(\mathrm{I}-\mathrm{H}_\lambda))} = MSE \quad (16)$$

where $tr(\mathrm{I} - \mathrm{H}_\lambda)^2$ represents the degrees of freedom (DF) for residuals. The quantity MSE is called the mean square Terror. As in linear regression, DF can be used in estimation of σ^2. Since MSE has a negligible bias term, the Eq. 16 is an unbiased estimate of σ^2 (Ruppert et al., 2003).

Furthermore, it is easy to see that the MSE which is the expected value of RSS is

$$E(RSS) = \sigma^2\left[n - tr\left(2\mathrm{H}_\lambda - \mathrm{H}_\lambda^2\right)\right] + E(\mathrm{T}_{\hat{G}})(\mathrm{I} - \mathrm{H}_\lambda)'(\mathrm{I} - \mathrm{H}_\lambda)E(\mathrm{T}_{\hat{G}}) = MSE$$

where the first term measures the variance, while the second term measures bias. To see these measures of the estimator, we expand $\widehat{\mathrm{B}}$ in (14b) by (12) to find

$$\widehat{\mathrm{B}} = (\mathrm{X}'\mathrm{U}^{-1}\mathrm{X})^{-1}\mathrm{X}'\mathrm{U}^{-1}\mathrm{T}_{\hat{G}} = \mathrm{B} + (\mathrm{X}'\mathrm{U}^{-1}\mathrm{X})^{-1}\mathrm{X}'\mathrm{U}^{-1}\mathrm{A}\hat{\mathrm{b}} + (\mathrm{X}'\mathrm{U}^{-1}\mathrm{X})^{-1}\mathrm{X}'\mathrm{U}^{-1}\varepsilon_{\hat{G}}$$

Hence, the bias and variance-covariance matrix of this estimator are, respectively,

$$Bias\left(\widehat{\mathrm{B}}\right) = E(\mathrm{B}) - \mathrm{B} = (\mathrm{X}'\mathrm{U}^{-1}\mathrm{X})^{-1}\mathrm{X}'\mathrm{U}^{-1}\mathrm{A}\hat{\mathrm{b}}$$
$$Var\left(\widehat{\mathrm{B}}\right) = \sigma^2(\mathrm{X}'\mathrm{U}^{-1}\mathrm{X})^{-1}\mathrm{X}'\mathrm{U}^{-1}\mathrm{X}'\mathrm{U}^{-1}(\mathrm{X}'\mathrm{U}^{-1}\mathrm{X})^{-1}$$

where σ^2 is illustrated in (16).

3. Choice of the smoothing parameter

In literature there are many classical methods used to select the amount of smoothing. Here we consider and compare the most widely used three selection criteria. As described in previous sections, these are AICc, GCV and Mallow's Cp. The positive parameter λ that minimizes any selection criteria is selected as an optimum smoothing parameter.

AICc criterion

Hurvich et al. (1998) proposed an improved version of classical Akaike's information criterion (AIC) to handle parameter selection problems. This selection criterion is denoted by AICc and defined as follows

$$AIC_c(\lambda) = 1 + \log[\|(\mathrm{H}_\lambda - \mathrm{I})\mathrm{T}_{\hat{G}}\|^2/n\,] + [2(tr(\mathrm{H}_\lambda) + 1)/n - tr(\mathrm{H}_\lambda) - 2]$$

where H_λ is a smoother matrix in (14c).

GCV criterion

The GCV is the most widely used criterion for selecting the smoothing parameter. It provides a more efficient calculation of leave-one out cross validation (CV). This criterion is described by Craven and Wahba (1978).

$$GCV(\lambda) = n^{-1}[\|(\mathrm{H}_\lambda - \mathrm{I})\mathrm{T}_{\hat{G}}\|^2]/[n^{-1}tr(\mathrm{I} - \mathrm{H}_\lambda]^2$$

Mallows' Cp criterion

The Cp is a common selection criterion proposed by Mallows (1973) for providing an MSE estimation scaled by σ^2. It is given as follows

$$C_p(\lambda) = \frac{1}{n}[\|(\mathrm{H}_\lambda - \mathrm{I})\mathrm{T}_{\hat{G}}\|^2 + 2\sigma^2 tr(\mathrm{H}_\lambda) - \sigma^2]$$

In practice if σ^2 is unknown, it is can be provided as

$$\hat{\sigma}^2_{\lambda_p} = \left\|\left(\mathrm{H}_{\lambda_p} - \mathrm{I}\right)\mathrm{T}_{\hat{G}}\right\|^2 / tr(\mathrm{I} - \mathrm{H}_{\lambda_p})$$

where λ_p is the pre-chosen value of the smoothing parameter with any selection method.

4. Numerical studies

In this section, we consider the different estimators for the components of the censored semiparametric regression model (12). It is discussed the estimator's finite properties with the two data sets. First data set is consisted by a real data from cancer patients, while second data set is obtained by a simulated data based on different censoring levels and sample sizes.

4.1. Real data example

We use a right censored real data from bowel cancer patients in Izmir, Turkey. To analyze the survival times of the patients we may write the semiparametric regression model corresponding to Eq. 11 as

$$L(Survival\ Time_i) = \beta_1 Age_i + \beta_2 Op.Time_i + \beta_3 Deg_i + f(Alb_i) + \varepsilon_i \quad (17)$$

where L (*Survival Time*) denote log survival times, *Age* and *Op.time* show age and operation time of the patients, respectively, *Deg* is degree of the cancer and *Alb* is albumin values from bowel cancer patients. The comparative outcomes from the parametric component of the model (17) are summarized in the Table 1. Note that the Table 1 show the estimated coefficients and their variances obtained by penalized spline estimators based on AIC, GCV and Cp criteria. From Table 1 we see that Cp has a better performance than other selection criteria in estimating the parametric component.

Table 1: The estimated regression coefficients and their variances from the model (16)

	β_1		β_2		β_3	
Criteria	Est.	Var.	Est.	Var.	Est.	Var.
AIC	-0.029	0.003	-0.005	0.001	0.266	0.114
GCV	-0.029	0.003	-0.005	0.001	0.261	0.110
Cp	-0.028	0.002	-0.005	0.001	0.277	0.102

To visualize function $f(.)$, we use a scatter diagram by plotting the smooth curve with 95% confidence intervals (see second panel of Fig. 1). Also, it is illustrated a scatter diagram of the survival times against *Alb* in the first panel of the Fig. 1. The second panel in Fig. 1 shows that relationship between Alb_i and $Survival\ Time_i$ may be nonlinear. In our context, the variable Alb_i is used as nonparametric part of the model (17). The main idea is to obtain a useful approximation $\hat{f}(Alb_i)$ to the real function $f(Alb_i)$. For these purposes, we fit a 3th-degree piecewise polynomial spline, as specified in (9), for the nonlinear $f(Alb_i)$using penalized spline method based on smoothing parameter λ selected by AICc, GCV, and Cp criteria, respectively. The estimated three spline fits are illustrated in the

Fig. 2. In order to assess the quality of these fits we used the mean square error (MSE) values, given by

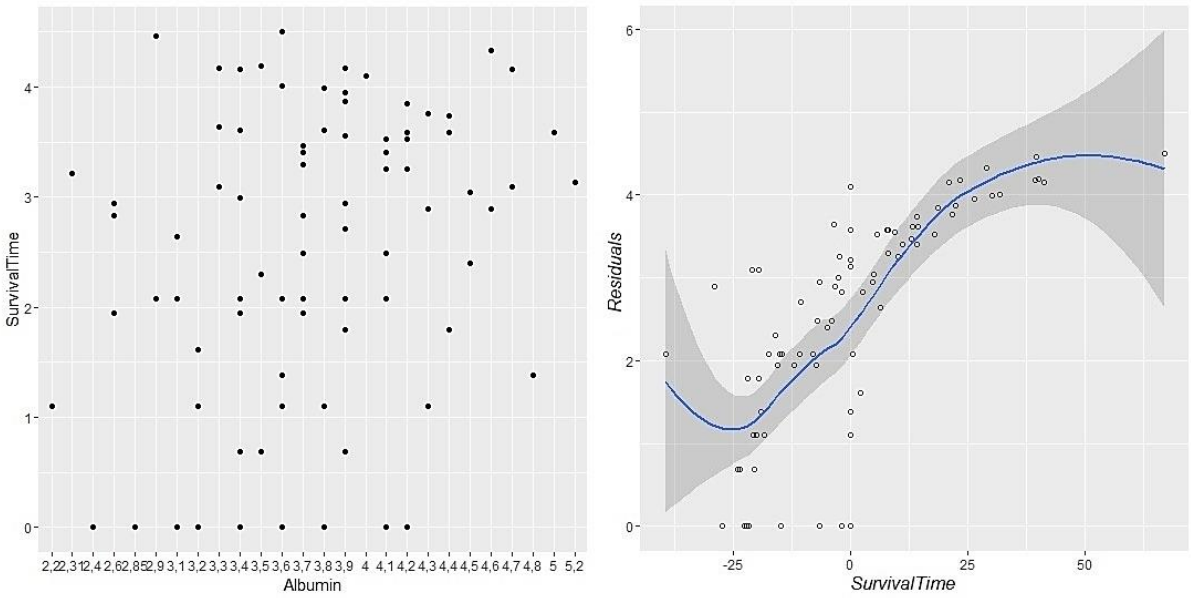

Fig. 1: Left panel denotes the scatterplot of Alb_i and $Survival\ Time_i$ whereas right panel shows the scatterplot residuals from regression of $Survival\ Time_i$ on Alb_i. The shaded regions denote the 95% confidence interval obtained by $\hat{f}(Alb_i) \pm 1.96\sqrt{Var\left(\hat{f}(Alb_i)\right)}$

$$MSE(\lambda) = \frac{1}{DF}\sum_{i=1}^{87}\left[f_\lambda(Alb_i) - \hat{f}_\lambda(Alb_i)\right]^2 \quad (18)$$

where $DF = n - tr(2\mathrm{H}_\lambda - \mathrm{H}_\lambda^2)$. The MSE values obtained from fits are 6.6779, 6.4497 and 5.9765, respectively.

It is easy to see that Cp outperforms the remaining two criteria, AICc and GCV, in estimating the nonparametric component of the model (17) with right censored data.

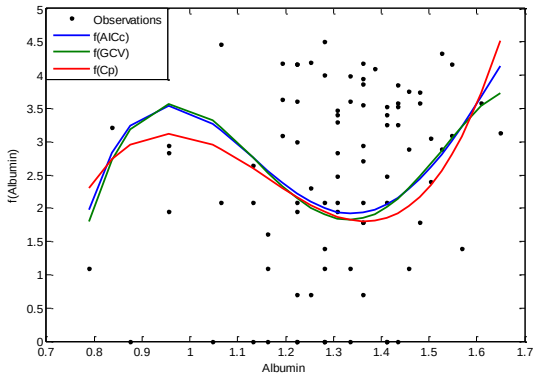

Fig. 2: Real observations and their smoothed curves obtained by penalized spline estimators based on AICc, GCV, and Cp criteria, respectively.

4.2. Simulation study

In this section, we carried out a Monte Carlo simulation study to show the practical performance of smoothing parameter criteria. To see the performance of the small, medium and large samples of each selection criteria, we consider three censoring levels (CLs), 10%, 30%, and 50% for each samples sizes with n = 50; 100; and 200. In this study, we generate the censoring variable as below;

Condition 1: $P(C) = |0.85 + 0.15\gamma|, if\ \gamma \le 1\ else\ P(C) = 0.90$
Condition 2: $P(C) = |0.65 + 0.15\gamma|, if\ \gamma \le 1\ else\ P(C) = 0.70$
Condition 2: $P(C) = |0.45 + 0.15\gamma|, if\ \gamma \le 1\ else\ P(C) = 0.50$

where $\gamma = |Z_i - 4|$ and $P(C)$ determines the probability of an observation satisfying $\gamma = 1$ (Wang et al., 2004). Under these conditions we obtain three censoring levels, 10%, 30% and 50%, respectively. The empirical data is generated by censored semiparametric regression model in generic form

$$T_i = 2X_{1i} + 1X_{2i} + 1.5X_{3i} + f(Z_i) + \varepsilon_i, i = 1, \dots, n \quad (19)$$

where X_j's have uniform distribution as $U(0,1)$, $(j=1,2,3)$, $f(Z_i) = \exp\{\sin(Z_i)\cos(3Z_i) + \sqrt{Z_i}\}$ with $Z_i = 4(i - 0.5)/n$and $Z_i \sim N(0,1)$. In here, however, because of the censoring, we consider transformed response (or synthetic) observations $T_{i\hat{G}}$ instead of T_i as described in the section 3.

For each censoring levels and sample sizes we conducted 1000 Monte Carlo simulation studies, and obtained 1000 estimates of the vector $\beta = (\beta_1, \beta_2, \beta_3)' = (2,1,-1.5)'$s forming the parametric part, and calculated 1000 smooth curves, constructing the nonparametric part, of the censored semiparametric model (19). The results of the simulation study are summarized in the rest of paper.

In this simulation study, because of nine different configurations are made, it is not possible to display all these configurations here. Therefore, only four different configurations are given in Figs. 1-4. The left panels of Figures show the estimates of $f(Z)$ obtained by selecting appropriate smoothing

parameters with AIC, GCV and Cp criteria under the mentioned censoring levels and sample sizes. The right panels of the same Figures display the box plots of the MSE values for nonparametric component estimates from the censored semiparametric model.

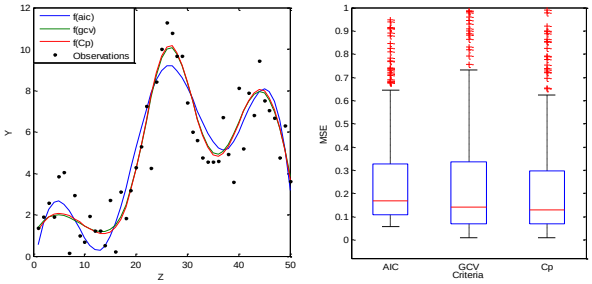

Fig. 3: The left panel displays three estimated curves of the nonparametric component together with real observations for n=50 and CL=10%. The right panel shows the box plots of the MSE values for AICc, GCV, and Cp criteria

Outcomes from the CL=30% for small sized samples are similar to Fig. 3 and are not reported. Furthermore, findings from the other sample sizes (n=100 and 200) for CL=10% and CL=30% (not displayed here) are similar to Fig. 3 and also not given here. Accordingly, it can be said that criteria are not superior to each other under the low and medium censoring levels. On the other hand, the effect of the heavy censoring levels is given in the Figs. 3-5 for small, medium, and large sample sizes, respectively.

It is seen from these Figures that the MSE values of the Cp are smaller than other criteria. This means that Cp is better than other criteria in terms of MSE for nonparametric component with heavy censoring level. Also, as can be seen Fig. 5, the two profile curves (corresponding to GCV and AICc) are located very close to each other.

In Table 2 "Bias" is biases of the estimated coefficients from simulation mean, "Std" is the simulation standard deviation, and "MSE" is the simulation mean square error of the estimated parameters, β_1, β_2 and β_3 respectively. Generally, the effect of the censoring levels tends to increase the variances of the estimated regression coefficients. The precision is declined as the censoring level increases.

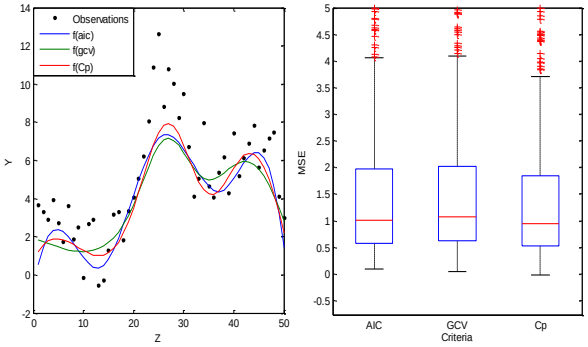

Fig. 4: Similar to Fig. 1 but for n=50 and CL=50%

Table 2: Simulation outcomes from parametric component of the censored semiparametric model (19)

	β_1			β_2			β_3		
	Bias	Std	MSE	Bias	Std	MSE	Bias	Std	MSE
				n=50, CL=10%					
AIC	0.368	0.226	0.187	0.191	0.174	0.036	0.319	0.206	0.127
GCV	0.398	0.223	0.236	0.205	0.166	0.028	0.837	0.203	0.136
Cp	0.346	0.221	0.126	0.179	0.165	0.027	0.797	0.199	0.113
				n=50, CL=30%					
AIC	0.396	0.342	0.064	0.130	0.328	0.055	0.008	0.309	0.040
GCV	0.458	0.347	0.072	0.143	0.331	0.049	0.172	0.309	0.023
Cp	0.372	0.321	0.056	0.172	0.318	0.043	0.354	0.299	0.0325
				n=50, CL=50%					
AIC	0.499	0.343	0.137	0.239	0.327	0.103	0.211	0.328	0.057
GCV	0.641	0.348	0.147	0.132	0.327	0.109	0.010	0.319	0.051
Cp	0.591	0.333	0.136	0.248	0.313	0.100	0.003	0.301	0.048
				n=200, CL=10%					
AIC	0.188	0.192	0.300	0.362	0.132	0.099	0.372	0.143	0.119
GCV	0.191	0.193	0.302	0.363	0.130	0.099	0.374	0.149	0.119
Cp	0.191	0.192	0.301	0.365	0.129	0.097	0.377	0.143	0.117
				n=200, CL=30%					
AIC	0.243	0.213	0.111	0.186	0.172	0.019	0.297	0.189	0.109
GCV	0.316	0.214	0.119	0.255	0.171	0.016	0.359	0.190	0.109
Cp	0.250	0.217	0.112	0.195	0.171	0.015	0.301	0.180	0.101
				n=200, CL=50%					
AIC	0.294	0.287	0.052	0.042	0.192	0.006	0.120	0.203	0.009
GCV	0.328	0.287	0.048	0.277	0.192	0.005	0.387	0.200	0.011
Cp	0.307	0.287	0.043	0.378	0.180	0.004	0.492	0.205	0.010

In addition, the precision is also improved as the sample size increases. To explain this issue, it is carried out the averages and the standard errors for the estimates of the regression coefficients obtained by the model (19) based on penalized spline for each criterion, sample, and censoring levels. The findings are illustrated in Table 2. In our context, Table 2 displays the biases and standard deviations of the vector $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)'$ over the 1000 simulations. To assess the quality of the vector $\hat{\beta}$, we used MSE value which is given by;

$$MSE_j = \frac{1}{1000}\sum_{i=1}^{1000}(\hat{\beta}_{ij} - \beta_j)^2, \quad j = 1,2,3 \qquad (20)$$

From the Table 1 we can see that the biases of $\hat{\beta}$ are very small. In this case, it can be said that

estimations of regression coefficients are quite satisfying for the three sample sizes and censoring levels. Especially the coefficients and their standard deviations estimated by Cp criterion are smaller than other criteria for all censoring levels and sample sizes. It appears that the Cp method generally outperforms other methods.

In addition, since the outcomes from medium sized samples (n = 100) are similar to those from large sized samples (n= 200), they are not reported here.

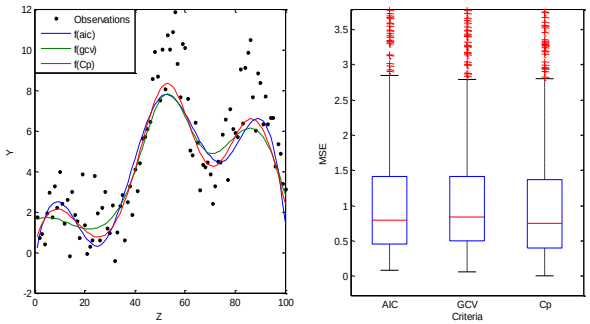

Fig. 5: Similar to Fig. 1 but for n=100 and CL=50%

In particular, when n is large (n = 200) which is illustrated in Fig. 6, the AICc method has small biases for all censoring levels. This means that the AICc method gives more unbiased estimates for large sized samples.

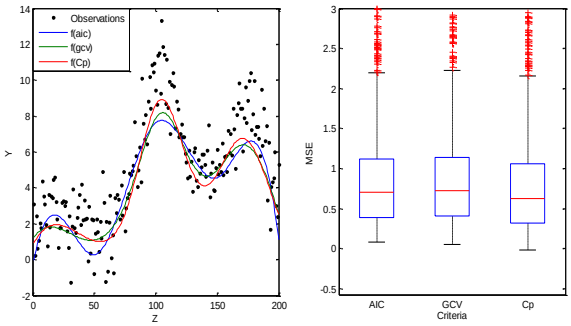

Fig. 6: Similar to Fig. 1 but for n=200 and CL=50%

However, when n is large, estimates of parametric coefficients, MSE, standard deviation values are more stable for Cp method under all censoring levels. Generally, the estimates obtained by three criteria are satisfying for all censoring levels. Furthermore, for each selection methods MSE values are very close to each other, especially for n = 200. This means that the remaining two estimators also work well. In summary, as can be seen from Table 2, the criteria giving smallest MSE are indicated by bold color. As expected, the MSE values are improved as the sample sizes increases.

Samples

There is not much of a difference between the three smoothing parameter selection criteria. However, MSE values of large sized samples are more stable than those of small sized samples in all censoring levels.

5. Conclusion

In this paper, we examined the performance of semiparametric model when response variable is randomly right-censored. As known, in semiparametric models, smoothing parameter selection plays an important role. Based on three selection criteria, we carried out three different models with right censored data. It is considered regression spline (penalized spline) method to estimate model parameters. In order to obtain accurate estimates, the smoothing parameter should be carefully chosen.

The main goal is to determine the selection criterion that gives better model fits. The following equation represents the MSE values of the fitted semiparametric models, estimated by averaging at the 1000 simulated data points. The mention MSE values are calculated as

$$MSE = \frac{1}{1000}\sum_{i=1}^{1000}\left(T_{i\hat{G}} - \hat{T}_{i\hat{G}}\right)^2 \qquad (21)$$

In this paper, we focused on the measure the performances of the selection methods and quality of the estimation method. For these purposes, we carried out both simulation study and real data example with using survival data. When we examine the results of simulation and real data experiments, we encountered some expected situations such as, big MSE values for high censoring levels for all selection methods, better estimations when low censoring levels or high sample sizes. In here Cp method illustrates better performance and estimations and note that AICc and GCV methods have similar results.

According to summarized results expressed in the Table 3, we can present the following recommendations and conclusions:

- Especially for large sample sizes and censoring levels, Cp has gave a better performance than AICc and GCV. In this case, the use Cp would be more beneficial.
- For small and medium sized samples, AIC gives a better performance than GCV and Cp criteria under the censoring level 30%. For especially medium censoring levels, the implementation of AIC would seem to be more appropriate.

Table 3: Estimated MSE values of the model for each selection methods

n	C.L.	AIC	GCV	Cp
50	10	0.388	0.388	0.386
	30	0.435	0.445	0.441
	50	0.515	0.525	0.509
100	10	0.208	0.209	0.208
	30	0.292	0.294	0.293
	50	0.338	0.342	0.333
200	10	0.139	0.139	0.131
	30	0.115	0.116	0.105
	50	0.261	0.262	0.252

Appendix A. Synthetic data transformation

As known, the mentioned regression function in this paper is a conditional expectation when response variable is censored from the right. Normally expectation of T_i values are

$$\frac{1}{n}\sum_{i=1}^{n} T_i$$

This is the standard estimation of $E(T_i)$. Because of censoring, instead of T_i, we use (L_i, δ_i) as in Eq. 2. If censoring distribution G is known, then the unbiased estimate of $E(T_i)$ will be as follows:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{L_i \delta_i}{G(L_i)}$$

In general, censoring distribution G is unknown. In this case, $E(T_i)$ on a certain s point is estimated with help of Kaplan-Meier mean

$$M_{KM} = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i L_i}{\hat{G}(L_i)} = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i T_i}{\hat{G}(T_i)} \quad = \int_0^{\infty} F(s|X=x, Z=z)dt \quad = \int_0^{\tau_F} F(s|X=x, Z=z)dt$$

Proof of equation is already illustrated by Susarla et al. (1984). According to that it can be said that;

$$M_{KM} \to E(T).$$

Appendix B. Synthetic response observation

When $n \to \infty$ it can be said that $E(\varepsilon_{iG}) \cong 0$ which is also means that $E(\varepsilon_i) \cong 0$. The vector form of this expression is given by $E(\varepsilon) \cong 0$. Let us consider the following model, $\mathrm{T} = \mathrm{XB} + \mathrm{Ab} + \varepsilon$ and $E(\mathrm{T}) = \mathrm{X}\hat{\mathrm{B}} + \mathrm{A}\hat{\mathrm{b}}$. Also $\hat{\mathrm{B}} = (\mathrm{X'U^{-1}X})^{-1}\mathrm{X'U^{-1}T}$ and $\hat{\mathrm{b}} = (\mathrm{A'A} + \lambda\mathrm{D})^{-1}\mathrm{A'}(\mathrm{T} - \mathrm{XB})$ it is easily seen that the model

$$\varepsilon = \mathrm{T} - \mathrm{X}\hat{\mathrm{B}} + \mathrm{A}\hat{\mathrm{b}} \quad = \mathrm{T} - \hat{\mathrm{T}}$$

The expected value of the ε is

$$E(\varepsilon) = E\left(\mathrm{T} - \hat{T}\right) = E(T) - E\left(\hat{T}\right) = \mathrm{X}\hat{\mathrm{B}} + \mathrm{A}\hat{\mathrm{b}} - E\left(\hat{\mathrm{T}}\right) = \mathrm{X}\hat{\mathrm{B}} + \mathrm{A}\hat{\mathrm{b}} - E\left(\mathrm{X}\hat{\mathrm{B}} + \mathrm{A}\hat{\mathrm{b}}\right) = \mathrm{X}\hat{\mathrm{B}} + \mathrm{A}\hat{\mathrm{b}} -$$
$$E\begin{pmatrix} \mathrm{X}(\mathrm{X'U^{-1}X})^{-1}\mathrm{X'U^{-1}T} \\ -\left((\mathrm{A'A} + \lambda\mathrm{D})^{-1}\mathrm{A'}\left(\mathrm{T} - \mathrm{X}\hat{\mathrm{B}}\right)\right) \end{pmatrix} = \mathrm{X}\hat{\mathrm{B}} + \mathrm{A}\hat{\mathrm{b}} -$$
$$E\begin{pmatrix} \mathrm{X}(\mathrm{X'U^{-1}X})^{-1}\mathrm{X'U^{-1}T} - (\mathrm{A'A} + \lambda\mathrm{D})^{-1}\mathrm{A'T} \\ -(\mathrm{A'A} + \lambda\mathrm{D})^{-1}\mathrm{A'X}(\mathrm{X'U^{-1}X})^{-1}\mathrm{X'U^{-1}T} \end{pmatrix} = \mathrm{X}\hat{\mathrm{B}} + \mathrm{A}\hat{\mathrm{b}} -$$
$$E(T - (\mathrm{A'A} + \lambda\mathrm{D})^{-1}\mathrm{A'T} + (\mathrm{A'A} + \lambda\mathrm{D})^{-1}\mathrm{A'T}) = \mathrm{X}\hat{\mathrm{B}} + \mathrm{A}\hat{\mathrm{b}} -$$
$$E(\mathrm{T}) = \mathrm{X}\hat{\mathrm{B}} + \mathrm{A}\hat{\mathrm{b}} - \mathrm{X}\hat{\mathrm{B}} - \mathrm{A}\hat{\mathrm{b}} = 0;$$

hence, as claimed for $n \to \infty$, $E(\varepsilon_{\mathrm{G}}) = E(\varepsilon) = 0$.

References

Brumback BA, Ruppert D, and Wand MP (1999). Comment. Journal of the American Statistical Association, 94(447), 794-797.

Buckley J and James I (1979). Linear regression with censored data. Biometrika, 66(3):429-436.

Chen S and Khan S (2001). Semiparametric estimation of a partially linear censored regression model. Econometric Theory, 17(3), 567-590.

Craven P and Wahba G (1978). Smoothing noisy data with spline functions. Numerische Mathematik, 31(4), 377-403.

Eilers PHC and Marx BD (2010). Splines, knots and penalties. Computational Statistics, 2(6):637- 653.

Hall P and Opsomer JD (2005). Theory for penalized spline regression. Biometrika, 92(1):105-118.

Hurvich CM, Simonoff JS, and Tasi CL (1998). Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. Journal of the Royal Statistical Society Series B, 60(2): 271-293.

Kaplan EL and Meier P (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53(282):457-481.

Koul H, Susarla V, and Van Ryzin J (1981). Regression analysis with randomly right-censored data. The Annals of Statistics, 9(6): 1276-1285.

Lai TL and Ying Z (1992). Asymptotically efficient estimation in censored and truncated regression models. Statistica Sinica, 2: 17-46.

Liang H (2006). Estimation in partially linear models and numerical comparisons. Computational statistics and data analysis, 50(3): 675-687.

Lu X and Cheng TL (2007). Randomly censored partially linear single-index models. Journal of Multivariate Analysis, 98(10): 1895-1922.

Mallows C (1973). Some comments on cp. Technometrics, 15(4): 661-675.

Miller RG (1976). Least squares regression with censored data. Biometrika, 63(3): 449-464.

Orbe J, Ferreira E, and Núñez-Antón V (2003). Censored partial regression. Biostatistics, 4(1): 109-121.

Qin G and Jing BY (2000). Asymptotic properties for estimation of partial linear models with censored data. Journal of Statistical Planning and Inference, 84(1): 95-110.

Ruppert D (2002). Selecting the number of knots for penalized splines. Journal of the Computational and Graphical Statistics, 11(4): 735-757.

Ruppert D, Wand MP, and Carroll RJ (2003). Semiparametric regression. Cambridge University Press. Cambridge, UK.

Stute W (1999). Nonlinear censored regression. Statistica Sinica, 9: 1089-1102.

Susarla V, Tsai WY, and Van Ryzin J (1984). A buckley-james-type estimator for the mean with censored data. Biometrika, 71(3): 624-625.

Wang Q, Linton O, and Hardle W (2004). Semiparametric regression analysis with missing response at random. Journal of the American Statistical Association, 99(466): 334-345.

Wang QH and Li G (2002). Empirical likelihood semiparametric regression analysis under random censorship. Journal of Multivariate Analysis, 83(2): 469-486.

Yu Y and Ruppert D. (2002). Penalized spline estimation for partially linear single-index models. Journal of the American Statistical Association, 97(460):1042-1054.

Zhou Y and Liang H (2009). Statistical inference for semiparametric varying-coefficient partially linear models with error-prone linear covariates. The Annals of Statistics, 37(1), 427-458.