Smoothing parameter selection in semiparametric regression models with censored data

Dursun Aydin*, Ersin Yilmaz

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

A R T I C L E  I N F O
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

A B S T R A C T
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.

© 2017 The Authors. Published by IASE. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

1. Introduction

Consider the semiparametric regression model (Eq. 1):

\[ T_i = X_i \beta_i + f(Z_i) + \varepsilon_i, i = 1, \ldots, n \]

(1)

where \( T_i \)'s are response observations, \( X_i = (X_{i1}, \ldots, X_{ip})\) is a \( p \)-dimensional vector, \( Z_i \)'s are observations of an extra univariate variable, \( \beta = (\beta_1, \ldots, \beta_p) \) is an unknown \( p \)-dimensional parameter vector to be estimated, \( f(\cdot) \) is an unknown smooth function and \( \varepsilon_i \)'s are random error terms with zero mean and variance \( \sigma^2 \). There are many approaches to estimate \( \beta \) 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 \( \beta \) and unknown function \( f \) in model (1), when the \( T_i \)'s are observed incompletely and right censored by a random variable \( C_i, i = 1, \ldots, n \) while \( X_i \) and \( Z_i \) are observed completely. Therefore, incomplete observations \( \{X_i, Z_i, T_i\} \) are transformed to complete observations \( \{X_i, Z_i, L_i, \delta_i, 1 \leq i \leq n\} \) as follows

\[ L_i = \min(T_i, C_i) \]

and

\[ \delta_i = I(T_i \leq C_i), i = 1, \ldots, n \]

(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 \( \delta_i \)'s are the values of the censoring indicator. As can be seen from the Eq. 2, censoring indicator provides the censoring information; if \( i \)th observation is complete then \( \delta_i \) 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
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

\[
F(s|X,Z) = P(T \leq s | X,Z), \quad G(s|X,Z) = P(C \leq s | X,Z)
\]

and

\[
M(s|X,Z) = P(L \leq s | X,Z) \quad (s \in R)
\]

from Eq. 3 their unknown survival functions are

\[
\tilde{F}(s|X,Z) = P(T > s | X,Z), \quad \tilde{G}(s|X,Z) = P(C > s | X,Z)
\]

and

\[
\tilde{M}(s|X,Z) = P(L > s | X,Z) \quad (s \in R)
\]

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

\[
\tilde{M}(s|X,Z) = \int \tilde{F}(s|X,Z) \tilde{G}(s|X,Z) dC(s|X,Z)
\]

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

\[
\tau_F = \inf\{s: \tilde{F}(s|X,Z) = 1\}, \quad \tau_G = \inf\{s: \tilde{G}(s|X,Z) = 1\}
\]

\[
\tau_M = \inf\{s: \tilde{M}(s|X,Z) = 1\}
\]

where \( \tau_F, \tau_G, \) and \( \tau_M \) are the common distribution of the censoring variable \( C_1 \) 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 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 \( T_i|\alpha \) according to the rule

\[
T_i|\alpha = \frac{\delta_{L_i}}{\delta_{\alpha}} = \frac{\delta_{L_i}}{1-\delta_{\alpha}}
\]

where \( G(.) \) is the common distribution of the censoring variable \( C_1 \) 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 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 \( T_i|\alpha \) according to the rule

\[
T_i|\alpha = \frac{\delta_{L_i}}{\delta_{\alpha}} = \frac{\delta_{L_i}}{1-\delta_{\alpha}}
\]

where \( G(.) \) is the common distribution of the censoring variable \( C_1 \) 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 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 \( T_i|\alpha \) according to the rule

\[
T_i|\alpha = \frac{\delta_{L_i}}{\delta_{\alpha}} = \frac{\delta_{L_i}}{1-\delta_{\alpha}}
\]
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, \( f(Z; α) \) is approximated by \( r \)th degree penalized spline with truncated polynomial basis (Ruppert et al., 2003):

\[
f(Z; α) = a_0 + a_1 Z + a_2 Z^2 + \cdots + a_r Z^r + \sum_{k=1}^K b_k (Z - \kappa_k)^{+}
\]

(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

\[
T_{ig} = \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + a_0 + a_1 Z_i + \cdots + a_r Z_i^r + \sum_{k=1}^K b_k (Z_i - \kappa_k)\]^{+} + \varepsilon_{ig}
\]

(11)

where \( i = 1, \ldots, n \) and \( T_{ig} \)'s are the synthetic response observations \( \varepsilon_{ig} \)'s are random error terms for given \( G \) and \( n \to \infty, E(\varepsilon_{ig}) = 0 \) (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)

\[
T_{gi} = XB + Ab + \varepsilon_g
\]

(12)

Where

\[
B = (\beta_1, \ldots, \beta_p, a_0, \ldots, a_r)\] \( B = (b_{1i}, \ldots, b_{ki})\), \( T_{gi} = (t_{g1}, \ldots, t_{gn}) \)

and

\[
\varepsilon_g = (\varepsilon_{g1}, \ldots, \varepsilon_{gn})\).

X and A are design matrices, given by

\[
X = \begin{bmatrix} X_{11} & \cdots & X_{1p} & 1 & \cdots & Z_i^r \end{bmatrix}, \ A = \begin{bmatrix} X_{n1} & \cdots & X_{np} & 1 & \cdots & Z_i^r \end{bmatrix}
\]

\[
\begin{bmatrix} (Z_1 - \kappa_1)^+ & \cdots & (Z_1 - \kappa_K)^+ \end{bmatrix}, \ \begin{bmatrix} (Z_n - \kappa_1)^+ & \cdots & (Z_n - \kappa_K)^+ \end{bmatrix}
\]

The penalized spline estimators \( \hat{B} = (\hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{a}_0, \ldots, \hat{a}_r) \), \( \hat{B} = (\hat{b}_{1i}, \ldots, \hat{b}_{ki}) \) of \( (B, b) \) are obtained by minimizing the penalized least-squares criterion

\[
PRSS(B; b) = (T_{gi} - XB - Ab)'(T_{gi} - XB - Ab) + \lambda b'Db
\]

(13)

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

The part \( \lambda b'Db \) 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 \( \lambda \) produce smoother estimators while smaller values produce more wiggly estimators. As can be seen from here, the parameter \( \lambda \) plays a key role in estimating of the model (11). Also, one of our tasks is to select an optimal value of the \( \lambda \) in here. This problem is discussed in section (4).

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

\[
[X'X \ X'A \ [A'X \ (A'A + \lambda D)]][B] = [X' \ 1']\hat{T}_G
\]

(14)

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

\[
b = (A'A + \lambda D)^{-1} A'(T_{\hat{G}} - XB)
\]

(14a)

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

\[
B = (X'U^{-1}X)^{-1}X'U^{-1}T_{\hat{G}}
\]

(14b)

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

\[
\hat{T}_G = (XB + \hat{b}) = H_bT_G
\]

where

\[
H_b = A(A'A + \lambda D)^{-1}A' + (I - A(A'A + \lambda D)^{-1}A)X(X'U^{-1}X)^{-1}X'U^{-1}
\]

(14c)

2.2. Estimation of variance

In practice smoothing parameter \( \lambda \) depends on variance of the error terms \( \varepsilon_{gi} = T_{gi} - Ab - XB \). As in linear regression, we can develop an estimator for \( \sigma^2 \) from the error or residuals sum of squares (RSS)

\[
RSS = \|T_{gi} - \hat{T}_G\|^2 = \|T_{gi} - H_bT_G\|^2 = \|I - H_b\|T_G\|^2
\]

(15)

where \( H_b \) is a smoother matrix, as defined in (14c). Thus, an estimator of \( \sigma^2 \) is obtained as

\[
\hat{\sigma}^2 = \frac{RSS}{tr(I - H_b)^2} = \frac{\|I - H_b\|T_G\|^2}{tr((I - H_b)(I - H_b))} = MSE
\]

(16)

where \( tr((I - H_b)^2) \) 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 \( \sigma^2 \). Since MSE has a negligible bias term, the Eq. 16 is an unbiased estimate of \( \sigma^2 \) (Ruppert et al., 2003).
Furthermore, it is easy to see that the MSE which is the expected value of RSS is
\[ E(\text{RSS}) = \sigma^2[n - \text{tr}(2H_\lambda - H_\lambda^2)] + E(T_\sigma)(1 - H_\lambda)\text{tr}(1 - H_\lambda) \]
where the first term measures the variance, while the second term measures bias. To see these measures of the estimator, we expand \( \hat{\beta} \) in (14b) by (12) to find
\[ \hat{\beta} = (X'U^{-1}X)^{-1}X'U^{-1}T_\sigma = B + (X'U^{-1}X)^{-1}X'U^{-1}\hat{\alpha} + (X'U^{-1}X)^{-1}X'U^{-1}\epsilon_\sigma \]

Hence, the bias and variance-covariance matrix of this estimator are, respectively,
\[ \text{Bias}(\hat{\beta}) = E(B) - B = (X'U^{-1}X)^{-1}X'U^{-1}\hat{\alpha} \]
\[ \text{Var}(\hat{\beta}) = \sigma^2(X'U^{-1}X)^{-1}X'U^{-1}(X'U^{-1}X)^{-1} \]
where \( \sigma^2 \) 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 Mallows’s Cp. The positive parameter \( \lambda \) 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
\[ AICc(\lambda) = 1 + \log[(H_\lambda - I)T_\sigma/n] + [2\text{tr}(H_\lambda) + 1]/n - \text{tr}(I - H_\lambda) \]
where \( H_\lambda \) 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).
\[ GCV(\lambda) = n^{-1}[\text{tr}(H_\lambda - I)T_\sigma^2] + 2\sigma^2\text{tr}(H_\lambda) - \sigma^2 \]

**Mallows’ Cp criterion**

The Cp is a common selection criterion proposed by Mallows (1973) for providing an MSE estimation scaled by \( \sigma^2 \). It is given as follows
\[ Cp(\lambda) = \frac{1}{n}[\text{tr}(H_\lambda - I)T_\sigma^2 + 2\sigma^2\text{tr}(H_\lambda)] - \sigma^2 \]

In practice if \( \sigma^2 \) is unknown, it is can be provided as
\[ \delta^2_p = \left\| (H_{\lambda_p} - I)T_\sigma \right\|^2/\text{tr}(1 - H_{\lambda_p}) \]
where \( \lambda_p \) 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
\[ L(\text{Survival Time}) = \beta_1 \text{Age}_i + \beta_2 \text{Op.Time}_i + \beta_3 \text{Deg}_i + f(\text{Alb}_i) + \epsilon_i \]
where \( L(\text{Survival Time}) \) denote log survival times, \( \text{Age} \) and \( \text{Op.time} \) show age and operation time of the patients, respectively, \( \text{Deg} \) is degree of the cancer and \( \text{Alb} \) 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>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>-0.029</td>
<td>0.003</td>
<td>-0.005</td>
<td>0.001</td>
<td>0.266</td>
<td>0.114</td>
</tr>
<tr>
<td>GCV</td>
<td>-0.029</td>
<td>0.003</td>
<td>-0.005</td>
<td>0.001</td>
<td>0.261</td>
<td>0.110</td>
</tr>
<tr>
<td>Cp</td>
<td>-0.028</td>
<td>0.002</td>
<td>-0.005</td>
<td>0.001</td>
<td>0.277</td>
<td>0.102</td>
</tr>
</tbody>
</table>

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 \( \text{Alb} \) in the first panel of the Fig. 1. The second panel in Fig. 1 shows that relationship between \( \text{Alb} \) and \( \text{Survival Time} \) may be nonlinear. In our context, the variable \( \text{Alb} \) is used as nonparametric part of the model (17). The main idea is to obtain a useful approximation \( f(\text{Alb}) \) to the real function \( f(\text{Alb}) \). For these purposes, we fit a 3rd-degree piecewise polynomial spline, as specified in (9), for the nonlinear \( f(\text{Alb}) \) using penalized spline method based on smoothing parameter \( \lambda \) selected by AICc, GCV, and Cp criteria, respectively. The estimated three spline fits are illustrated in the
used the mean square error (MSE) values, given by

$$\text{MSE}(\lambda) = \frac{1}{\text{DF}} \sum_{i=1}^{87} \left[f_i(\text{Alb}_i) - \hat{f}_i(\text{Alb}_i)\right]^2$$

where $\text{DF} = n - \text{tr}(2\mathbf{A} - \mathbf{H}^2)$. 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.

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

$$\text{MSE}(\lambda) = \frac{1}{\text{DF}} \sum_{i=1}^{87} \left[f_i(\text{Alb}_i) - \hat{f}_i(\text{Alb}_i)\right]^2$$

where $\text{DF} = n - \text{tr}(2\mathbf{A} - \mathbf{H}^2)$. 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.

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:

$$T_i = 2X_{i1} + 1X_{i2} + 1.5X_{i3} + f(Z_i) + \varepsilon_i, i = 1, ..., n$$

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 $T_{i|\hat{c}}$ 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.

![Figure 3](image)

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

Outcomes from the \( \text{CL}=30\% \) 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 \( \text{CL}=10\% \) and \( \text{CL}=30\% \) (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.

![Figure 4](image)

**Fig. 4:** Similar to Fig. 1 but for \( n=50 \) and \( \text{CL}=50\% \)

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;

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

(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 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, \( \hat{\beta}_1, \hat{\beta}_2 \) and \( \hat{\beta}_3 \) 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.

![Table 2](image)

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

The left panel displays three estimated curves of the nonparametric component together with real observations for \( n=50 \) and \( \text{CL}=10\% \). The right panel shows the box plots of the MSE values for AICc, GCV, and Cp criteria.
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.

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.

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

\[
MSE = \frac{1}{1000} \sum_{i=1}^{1000} (\hat{T}_i - \tilde{T}_i)^2
\]

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

<table>
<thead>
<tr>
<th>n</th>
<th>C.L.</th>
<th>AIC</th>
<th>GCV</th>
<th>Cp</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>30</td>
<td>0.388</td>
<td>0.388</td>
<td>0.386</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.415</td>
<td>0.445</td>
<td>0.441</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.208</td>
<td>0.209</td>
<td>0.208</td>
</tr>
<tr>
<td>100</td>
<td>30</td>
<td>0.292</td>
<td>0.294</td>
<td>0.293</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.338</td>
<td>0.342</td>
<td>0.333</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.139</td>
<td>0.139</td>
<td>0.131</td>
</tr>
<tr>
<td>200</td>
<td>30</td>
<td>0.115</td>
<td>0.116</td>
<td>0.105</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.261</td>
<td>0.262</td>
<td>0.252</td>
</tr>
</tbody>
</table>
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

\[
\sum_{i=1}^{n} \frac{1}{n} T_i
\]

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

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

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

\[
M_{KM} = \sum_{i=1}^{n} \frac{L_i \delta_i}{n \bar{G}(L_i)} = \int_0^\infty F(s|X = x, Z = z) dt = \int_0^\infty F(s|x = x, Z = z) dt
\]

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

\[
M_{KM} \rightarrow E(T).
\]

Appendix B. Synthetic response observation

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

\[
\varepsilon = T - X\hat{B} + A\hat{b} = T - \hat{T}
\]

The expected value of the $\varepsilon$ is

\[
E(\varepsilon) = E(T - \hat{T}) = E(T) - E(\hat{T}) = X\hat{B} + A\hat{b} - E(\hat{T}) = X\hat{B} + A\hat{b} - \frac{X(X'U^{-1}X)^{-1}X'U^{-1}T}{(A'\hat{A} + \lambda D)^{-1}A'(T - X\hat{B})} = X\hat{B} + A\hat{b} - E(T - (A'\hat{A} + \lambda D)^{-1}A'T - (A'\hat{A} + \lambda D)^{-1}A'T) = X\hat{B} + A\hat{b} - E(T) = X\hat{B} + A\hat{b} - X\hat{B} - A\hat{b} = 0;
\]

hence, as claimed for $n \rightarrow \infty$, $E(\varepsilon_i) = E(\varepsilon) = 0$.

References


