Right-Censored Nonparametric Regression: A Comparative Simulation Study

Dursun Aydın [1], Ersin Yılmaz [1]

[1] *Mugla Sitki Kocman University, Faculty of Science, Department of Statistics, Mugla, Turkey*

***Abstract* – This paper introduces the operating of the selection criteria for right-censored nonparametric regression using smoothing spline. In order to transform the response variable into a variable that contains the right-censorship, we used the Kaplan-Meier weights proposed by [1], and [2]. The major problem in smoothing spline method is to determine a smoothing parameter to obtain nonparametric estimates of the regression function. In this study, the mentioned parameter is chosen based on censored data by means of the criteria such as improved Akaike information criterion (AICc), Bayesian (or Schwarz) information criterion (BIC) and generalized cross-validation (GCV). For this purpose, a Monte-Carlo simulation study is carried out to illustrate which selection criterion gives the best estimation for censored data.**

***Keywords* – Nonparametric Regression, Spline Smoothing, Kaplan-Meier weights, Censored data.**

1. Introduction

Consider the following nonparametric regression model

$$Y = f(Z) + \varepsilon \qquad (1)$$

where Y is a response variable, $f(.)$ is an unknown smooth function, Z is a covariate, and ε is a random error term with mean zero and constant variance σ^2. Suppose that $\{Y_i, Z_i, i \le 1 \le n\}$ is a random sample satisfying the model (1). The model (1) was discussed by [3] based on the assumption of the completely observed Y. In our study, we are interested in estimating the unknown function $f(.)$ when Y is observed incompletely and right censored by a random variable C, but Z are observed completely. Therefore, instead of observing $\{Y_i, Z_i\}$ we observe $\{(Z_i, T_i, \delta_i), i \le 1 \le n\}$ with

$$T_i = \min(Y_i, C_i) \text{ and } \delta_i = \mathrm{I}(Y_i \le C_i), \quad i \le 1 \le n \qquad (2)$$

where T_i and C_i are referred to as the lifetimes, the censoring time, respectively, for the ith subject. Y_i is the observed lifetimes, while the censoring indicator $\delta_i = \mathrm{I}(.)$ stores up the information whether an observation is censored or uncensored. Also, we assume that the C_i's are independently distributed as some unknown censoring distribution G, and Y_i and C_i are independent.

The problem of censored regression is now to estimate the unknown regression function $f(Z) = E(T|Z = z)$ from the data (Z, T, δ). In this case, the relationship between the updated response variable T and Z are projected by

$$T_i = f(Z_i) + \varepsilon_i \qquad (3)$$

where ε_i's are the random error terms independent of Z, and T_i's are defined as in (2). Then, the smoothing spline estimates of the function $f(.)$ are obtained by minimizing the weighted penalized residuals' sum of squares

$$PRSS = \sum_{i=1}^{n} w_i \left(t_{(i)} - f(z_{(i)})\right)^2 + \lambda \int_a^b \left(f''(z)^2\right) dz \qquad (4)$$

where $f \in [a,b]$ is an unknown smooth function with continuous first and second derivatives, $t_{(i)}$'s are the ordered values of T, $z_{(i)}$'s are the ordered values of Z, which is the concomitant associated with T, the

DOI: 10.18421/TEM54-06
https://dx.doi.org/10.18421/TEM54-06

Corresponding author: Dursun Aydın,
Faculty of Science, Department of Statistics, Muğla, Turkey
Email: duaydin@mu.edu.tr

λ is a positive smoothing parameter, $\int f''(z)^2 dz$ is penalty term for smoothing spline based on $\lambda > 0$, and w_i's are the Kaplan-Meier (K-M) weights connected to $t_{(i)}$, and they are also the ith diagonal element of the matrix W with entries

$$w_i = \frac{\delta_{(i)}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1} \right)^{\delta_{(j)}} \tag{5}$$

$\delta_{(i)}$'s are the corresponding censoring indicators associated with $t_{(i)}$. Specifically, trace of the diagonal matrix W is equal to one.

Many authors have dealt with the estimation problem of the nonparametric regression model. Examples of this work include [4], [5], [6], [7], and [8]. For example [9] focuses on the spline estimates of a partially linear model. Some authors consider the estimation of the residual variance in nonparametric regression ([10]). Classical Akaike information criterion for the smoothing parameter selection in nonparametric regression is improved by [11]. Linear smoother and additive models are discussed by [12]. A nonparametric estimator of a regression function based on censored data is studied by [1]. Empirical likelihood semi parametric random censorship models are considered by [13]. Also, the other key references for regression with censored data are [14], and [2].

2. Weighted smoothing spline

Smoothing spline performs a regularized regression over the natural spline basis, placing knots at all points $\{z_i, i = 1, 2, ..., n\}$. This method uses the input points as knots and thus it overcomes the knot selection problem and simultaneously, smoothing spline avoids the overfitting by shrinking the coefficients of the estimated function.

The smoothing spline fits for the model (2) is obtained by solving the minimization problem, given in matrix and vector form

$$\left\| Y - N\mathbf{f} \right\|_2^2 + \lambda \mathbf{f}' K \mathbf{f} \tag{6}$$

where N is a $n \times q$ incidence matrix, with elements

$$N_{ij} = \mathrm{I}\left(z_i = s_j\right),\ 1 \le i \le n,\ 1 \le j \le q$$

which consist of the s_j values, distinct and ordered values of the knot points z_i, and K is a penalty matrix, given by $K = Q'R^{-1}Q$ (see [7] for more details).

The solution to minimization problem (6) is a natural cubic spline with knots s_j (see [7]). There is a matrix K only depending on the knots s_j, such that the minimized value of $\int f''(z)^2 dz$ in (4) equals to $\mathbf{f}' K \mathbf{f}$ in (6). That is, $\mathbf{f}' K \mathbf{f} = \int f''(z)^2 dz$, where the K is also a symmetric $n \times n$ positive definite penalty matrix with solution $\lambda K = \mathbf{S}_\lambda^{-1} - \mathrm{I}$. In this case, within optimal $\mathbf{f}$ minimizing the equation (6), smoothing spline curve is calculated as

$$\hat{\mathbf{f}}_\lambda = \left(\mathrm{I} + \lambda K\right)^{-1} Y = \mathbf{S}_\lambda Y \tag{7}$$

where λ is the smoothing parameter which adjusts the penalty term, as said before, and $\mathbf{S}_\lambda = \left(\mathrm{I} + \lambda K\right)^{-1}$, is a well-known positive-definite spline smoother matrix which depends on λ.

In this work we propose a smoothing spline method to fit model (1) when the dependent variable Y is at risk of being censored. For this reason, the smoothing spline method for estimating $f(.)$ can not be applied directly here. To overcome this problem we used the weighted smoothing spline method that is discussed by [15].

As we have mentioned above, because of the censoring, instead of observing $\{Y_i, Z_i\}$ we observe $\{Z_i, T_i, \delta_i\}$. Thus, the response observations are updated as T_i via the equation (2). Then, the spline that fits for the model (3) is carried out by solving the minimization problem (4).

We can represent the equation (4) in matrix and vector form

$$PRSS = \left(T - \mathbf{f}\right)' W \left(T - \mathbf{f}\right) + \lambda \mathbf{f}' K \mathbf{f} \tag{8}$$

where W is a diagonal matrix formed with K-M weights w_i in (5). In its simplest form (8) could be seen as a weighted version of the equation (6). In a similar manner to (7) the weighted smoothing spline fits for the model (3) are obtained by

$$\hat{\mathbf{f}}_\lambda = \left(W + \lambda K\right)^{-1} W T \tag{9}$$

As expressed before, the most important issue in this method is to select the smoothing parameter λ For this purpose, it is considered the most widely used three criteria, given in the next section.

3. Selection criteria

The positive value λ that minimizes any smoothing parameter selection methods is selected as an appropriate smoothing parameter.

Akaike information criterion (AICc): An improved version based on the classical Akaike criterion is developed by [11]:

$$AIC_c = 1 + \log\left(\left\|(\mathbf{H}_\lambda - \mathrm{I})T\right\|^2 \big/ n\right) + \left[\left(2tr\left(\mathbf{H}_\lambda\right) + 1\right)\big/ n\right] - tr\left(\mathbf{H}_\lambda\right) - 2 \qquad (10)$$

where $\mathbf{H}_\lambda = \left(W + \lambda K\right)^{-1} W$ is a hat matrix that plays a similar role to spline smoother matrix in (7).

B*ayesian information criterion (BIC):* [16] improved by using Bayes estimators. The generic form of the BIC criterion is

$$\mathrm{BIC} = \frac{1}{n}\left\|\left(I\text{-}\mathbf{H}_\lambda\right)T\right\|^2 + \frac{\log(n)}{n} tr\left(\mathbf{H}_\lambda\right) \qquad (11)$$

***Genralized Cross-Validation (GCV) criterion*:** The criterion function is defined by [4], and described as

$$\mathrm{GCV} = n^{-1}\left\|\left(I\text{-}\mathbf{H}_\lambda\right)T\right\|^2 \Big/ \left[n^{-1} tr\left(I - \mathbf{H}_\lambda\right)\right]^2 \qquad (12)$$

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

4. Estimating the variance

The main goal is to select an appropriate estimator of f from among the elements $\left[\hat{f}_\lambda : \lambda \in R \,\middle|\, \lambda > 0\right]$. In order to find an optimum estimator there are some performance measures which are widely used and accepted. The mean square error (MSE) of prediction, one of these measures can be obtained by average value of residuals sum of squares $n^{-1}RSS\left(\lambda\right)$. The mentioned residual sum of squares (*RSS*) is defined as

$$RSS\left(\lambda\right) = \sum_{i=1}^{n}\left((\hat{f}_\lambda)_i - T_i\right)^2 \qquad (13)$$

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

$$\begin{aligned} RSS\left(\lambda\right) &= \left(\hat{\mathbf{f}}_\lambda - \mathbf{T}\right)'\left(\hat{\mathbf{f}}_\lambda - \mathbf{T}\right) \\ &= \mathbf{T}\left(\mathbf{I} - \mathbf{H}_\lambda\right)^2 \mathbf{T} \end{aligned} \qquad (14)$$

where $\hat{\mathbf{f}}_\lambda = \left(W + \lambda K\right)^{-1} W T = \mathbf{H}_\lambda \mathbf{T}$ is defined as in (9). The expected value of squared residuals given in (13) or (14) is also known as *MSE* of prediction, which in this case is

$$\begin{aligned} MSE\left(\lambda\right) &= E\left\|\mathbf{T} - \hat{\mathbf{f}}_\lambda\right\|^2 \\ &= E\left\|\left(\mathbf{I}\text{-}\mathbf{H}_\lambda\right)\mathbf{T}\right\|^2 \end{aligned} \qquad (15)$$

It follows directly from (15) that $MSE\left(\lambda\right)$ can be described as

$$\begin{aligned} MSE\left(\lambda\right) &= \mathbf{f}'_\lambda\left(\mathbf{I} - \mathbf{H}_\lambda\right)^2 \mathbf{f}_\lambda \\ &\quad + \sigma^2\left[n - 2\left(\mathbf{H}_\lambda\right) + \left(\mathbf{H}_\lambda{}'\mathbf{H}_\lambda\right)\right] \end{aligned} \qquad (16)$$

Hence, follows the equation (16) that $n^{-1}RSS\left(\lambda\right)$ is a biased estimator of MSE (see [8]).

In practice, the equation (16) cannot be computed directly because 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 (14).

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

$$\hat{\sigma}_\varepsilon^2 = \frac{RSS\left(\lambda\right)}{n - p} = \frac{RSS\left(\lambda\right)}{tr\left(\mathbf{I} - \mathbf{H}_\lambda\right)^2} = \frac{RSS\left(\lambda\right)}{DF_{RES}} \qquad (17)$$

where $RSS\left(\lambda\right)$ is defined as in (19) , and

$$\begin{aligned} DF_{RES} &= tr\left(\mathbf{I} - \mathbf{H}_\lambda\right)^2 \\ &= n - 2tr\left(\mathbf{H}_\lambda\right) + tr\left(\mathbf{H}_\lambda{}'\mathbf{H}_\lambda\right) \end{aligned} \qquad (18)$$

called the residual degrees of freedom (DF_{RES}) for pre- chosen λ 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 (18) is an unbiased estimate of σ_ε^2.

5. Simulation study

In this section, we performed a simulation study to assess the operating of the selection criteria introduced in Section 3. To see the performance of the small, medium and large samples of each criteria, we consider three censoring levels (CLs), 15%, 30%, and 45% and three samples sizes with n = 50, 100, and 200. The number of replication was 1000 for each of the samples. Our data is generated by censored nonparametric model in generic form

$$T_i = \left\{f\left(Z_i\right) = 1.2 z_i\left(\sin z_i\right)\right\} + \varepsilon_i, \quad \varepsilon_i \sim N\left(0, \sigma = 1\right)$$

where $z_i = 15\left[\left(i - 0.5\right)/n,\ i = 1, 2, ..., n\right]$.

Furthermore, we used the values of mean square error (MSE) to evaluate the quality of any curve estimate $(\hat{\mathbf{f}}_\lambda)$:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left\{ f(z_i) - \hat{f}_\lambda(z_i)\right\}^2 , \; \hat{f}_\lambda(z_i) = (\hat{\mathbf{f}}_\lambda)_i \qquad (17)$$

Monte Carlo simulation results are illustrated in the following Figures and Tables.

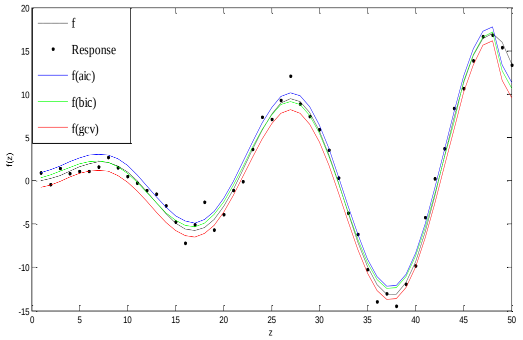

Figure 1: Real data and the true function together with its smooth curves based on AICc, BIC and GCV criteria for n=50, and CL=30%.

As can be seen from Figures 1-3, the estimated functions become closer to the real function when sample size increases, regardless of the levels of censoring. However, the estimated curve with BIC criterion does not work as favorably for small sized data sets.

Generally, the effect of the censoring tends to increase the variance of the estimators. The precision is declined as the censoring level increases. In addition, the precision is also improved as the sample size increases. To explain this issue, the MSE values in (13) are computed from the spline fits for each criterion, sample, and censoring levels. The findings are shown in Table 1.

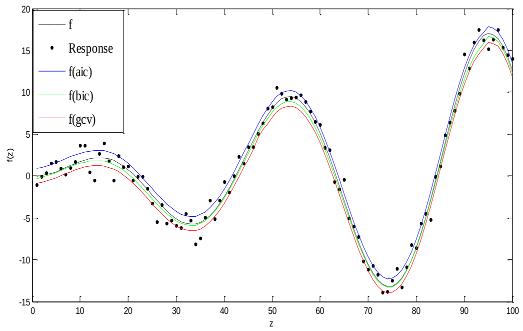

Figure 2: Similar to Figure 1, but for n=100, and CL=45%

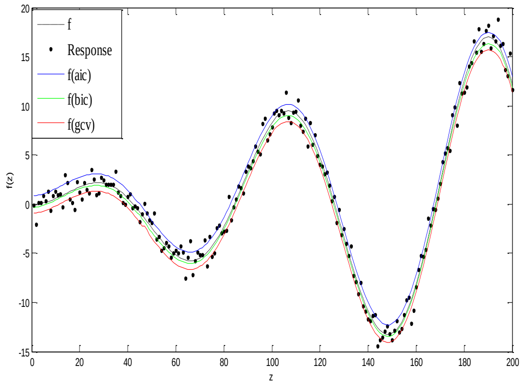

Figure 3: Similar to Figure 1, but for n=200, and CL=45%.

Table 1: MSE values for nonparametric models

n	CLs (%)	AICc	BIC	GCV
50	15	**0.1603**	0.1630	0.1625
	30	0.3572	0.3652	**0.3625**
	45	**0.5944**	0.6441	0.6228
100	15	0.1154	0.1131	**0.1152**
	30	**0.1814**	0.1827	0.1832
	45	**0.2811**	0.2888	0.2864
200	15	**0.1074**	0.1073	**0.1074**
	30	**0.1501**	**0.1501**	0.1502
	45	**0.1977**	0.1979	0.1979

As can be seen from Table 1., the criteria giving smallest MSE are indicated by bold color. As expected, the MSE values are improved as the sample sizes increases. From this, it is easily understood that AICc provides a good parameter λ in general.

Boxplots for MSE values based on each criterion are illustrated in Figure 4. In this Figure, A1, A2 and A3 denote the MSE values based on AICc for sample sizes n=50,100 and 200, respectively. In a similar fashion, B1, B2 and B3 show the MSE values for BIC. Finally, G1, G2 and G3 indicate the MSE values for GCV. Also, the upper panel of Figure 4. has CL=15%, medium panel CL=30%, and bottom panel CL= 45%.

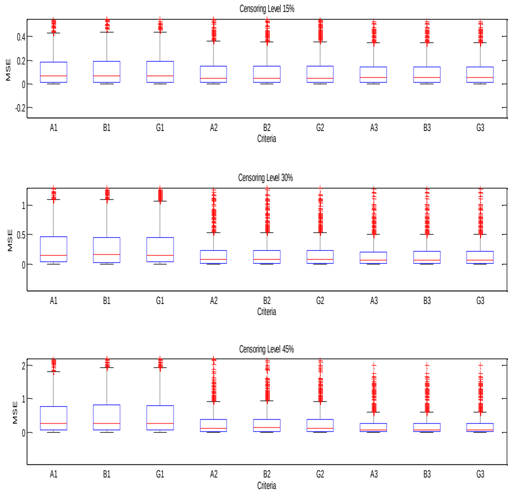

Figure 4: Boxplots of the MSE values for estimated nonparametric models

6. Concluding remarks

We have shown some useful results for selecting the smoothing parameter in the censored nonparametric regression models. The outcomes suggest that when we choose the parameter λ giving smallest AICc, the obtained spline estimators outperform the others in terms of MSE values for sample sizes with n=50 and 100. On the other hand, AICc criterion also produces good estimates for the right censored nonparametric models under different censoring levels.

The simulation experiment results are satisfactory in general. Also, as sample sizes increase, for each selection criterion the right censored nonparametric model indicates a closer fit to real observations. As a result, it can be said that there is no notable difference between selection methods in selection smoothing parameter for large sized samples. More specifically, the estimators based on AICc, GCV, and BIC gave the same MSE values for n=200, and all censoring levels. However, BIC criterion produced poor performance in this setting.

Finally, by considering the simulation results, we can propose the following key ideas:

- For especially small sized samples, AICc is recommended as being a good selection criterion.
- For large samples, the implementation of AICc in addition to BIC and GCV criteria would be more beneficial.

References

[1]. Stute, W. (1993), Consistent Estimation Under Random Censorship When Covariables are Present, Journal of Multivariate Analysis, Vol. 45, 89-103.

[2]. Orbe, J., Ferreira, E., Nunez-Anton, V. (2003), Censored Partial Regression, Biostatistics, Vol.4, No.1, 109-121.

[3]. Engle, R. F., Granger, C. W., Rice, J., Weiss, A. (1986), Semiparametric Estimates Of The Relation Between Weather and Electricity Sales, Journal Pf The American Statistical Association, 310-320.

[4]. Craven, P., Wahba, G. (1979), Smoothing Noisy Data with Spline Functions, Numeriche Mathematik, Vol. 31(4), 377-403.

[5]. Silverman, B.W. (1984), Spline smoothing: The Equivalent Variable Kernel Method, The Annals of Statistics, Vol. 12, No.3, 898-916.

[6]. Hardle, W. (1990), Applied Nonparametric Regression, Cambridge University Press.

[7]. Green, P. J., Silverman, B. W. (1994), Nonparametric Regression and Generalized Linear Models, Chapman & Hall, London.

[8]. Eubank, L. R., (2000), Spline Regression, Smoothing and Regression: The Annals of Statistics, Vol. 12, 1215-1230.

[9]. Heckman, N. E., (1986), Spline Smoothing in a Partly Linear Model, Journal of the Royal Statistical Society, Series B, Vol. 48(2), 244-248.

[10]. Buckley, M. J., Eagleson, G. K., Silverman, B. W. (1988), Estimation of Residual Variance in Nonparametric Regression, Biometrika, Vol. 75, 183-199.

[11]. Hurvich, C. M., Simonoff, J. S., Tasi, C. L. (1988), Smoothing Parameter Selection in Nonparametric Regression Using An Improved Akaike Information Criterion, J. R. Statist. Soc. B., Vol. 60, 271-293.

[12].Buja, A., Hastie, T. J., Tibshirani, R. J. (1989), Linear Smoother and Additive Models, The Annals of Statistics, Vol. 17, 81-89.

[13]. Wang, Q.-H., Li, G. (2002), Empirical Likelihood Semiparametric Regression Analysis Under Random Censorship, Journal Of Multivariate Analysis, Vol.83(2), 469-486.

[14]. Miller, R., Halpern, J. (1982), Regression With Censored Data, Biometrika, Vol. 69, 521-531.

[15].Rodrigez, G., (2001), Smoothing and Non-Parametric Regression, Spring.

[16]. Schwarz G., (1978), Estimation the dimension of a model, The Annals of Statistics, Vol. 6(2), 461-464.