Nonparametric Regression with Randomly Right-Censored Data

DURSUN AYDIN ERSİN YILMAZ
Faculty of Science, Department of Statistics
Mugla Sitki Kocman University
Muğla, Turkey, 48000
daydin@mu.edu.tr ersinyilmaz@mu.edu.tr

Abstract: - The purpose of this study is to estimate the right-censored nonparametric model with kernel smoothing method. To consider the censorship, we used Kaplan-Meier estimator proposed by Stute (1993). In nonparametric statistics, a kernel smoothing method needs a smoothing parameter which is also called as a bandwidth parameter. In this study, we choose the bandwidth parameter by using three selection methods such as improved version of Akaike information criterion (AICc), Risk estimation using classical pilots (RECP) and Generalized cross-validation(GCV) method, respectively. For this purpose, a Monte-Carlo simulation study is performed to illustrate which selection criterion gives the best estimation for different sample sizes and censoring levels.

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

1 Introduction and main ideas

In a simple manner, the concept of the censored data represents the incomplete observation data. It can be encountered with this term in many working such as health and survival analysis. In survival setting we can explain the censorship as if lifetime values measured from objects or subjects cannot be completely observed, then data censored. According to this we can obtain only partial information that could be considered as a censoring variable.

Generally, hazard risk methods and parametric regression approach are used for estimating censored data. Although these methods are very popular for censored data, their restrictions and assumptions are obstacles for the accuracy of the estimation and disadvantage for usage fields. In this study, we solve this problem with using nonparametric regression model that free from assumptions. The mentioned nonparametric regression model can be expressed as follows

$$Y_i = f(X_i) + \varepsilon_i,\ 1 \le i \le n, \qquad (1)$$

where Y_i's are the right-censored response values and X_i's represent the values of the nonparametric covariate variable and ε_i's are independent and identically distributed random errors with zero mean and constant variance σ^2 and $f(.)$ is a unknown smooth regression function.

As known, nonparametric and semi-parametric models are very popular recently. To this respect, we focus on the estimate unknown function in model (1) with the right-censored variable *Y*. Note that response variable *Y* is also censored form right by censoring variable C but *X* is observed completely.Therefore, instead of observing (*X*, *Y*) we observe the triplets (*X*, *T*, δ). In this case, we define a new adjusted response variable which includes the minimum values of *Y* and C. The new response observations are

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

where $\delta_i = \mathrm{I}(.)$ is censor indicator variable and it is a sign function. If response value is censored, there is an incomplete observation and $\delta_i = 0$, and $\delta_i = 1$ 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

A1. *C* is independent from *X* and *Y*.
A2. $\mathrm{P}(Y \le C|Y,X) = \mathrm{P}(Y \le C|Y)$

Also, assume that *Y* and *C* independent and non-negative variables and they have unknown distribution functions *F* and *G*, respectively. In here, A1 is the known censorship assumption when we use Kaplan Meier (1958) estimator. A2 means that, given the lifetime, whether there is a censorship or not, covariate variable do not ensure any more information.

We see many studies in literature about estimation of nonparametric model with kernel smoothing. Examples of this work include Wand et. al., (1984), Hardle (1990), Green and Silverman (1994), 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 semiparametric random censorship models are discussed by Wang and Li (2002).

This paper is organized as follows. Section 2 introduces the computations of kernel smoothing estimation within right-censored data. Bandwidth selection methods are given in section 3. A simulation study is carried out to compare the selection methods in section 5 and conclusions are discussed in section 5.

2 Kernel Smoothing Method

In this study, we propose a kernel smoothing method to fit model (1) when the dependent variable Y is at risk of being censored. For this reason, the traditional kernel smoothing method for estimating $f(.)$ 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 order to solve this problem Koul et al. (1981) proposed to replace G by its the Kaplan-Meier estimator:

$$\hat{G}(t)=1-\prod_{i=1}^{n}\left(\frac{n-i}{n-i+1}\right)^{\mathrm{I}[t_{(i)}\le t,\ \delta_{(i)}=0]},\ (t\ge 0) \qquad (3)$$

where $t_{(1)}\le ...\le t_{(n)}$ are ordered observations of *T*, and $\delta_{(1)}\le ....\le \delta_{(n)}$ are the corresponding censoring indicators observations (censored and uncensored lifetimes), which is the concomitant associated with *T*. Using the equation (3) the synthetic response variable can be obtained as

$$T_{i\hat{G}}=\left(\delta_i T_i\right)\Big/\left(1-\hat{G}(T_i)\right),\ \ i=1,2,...,n \qquad (4)$$

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

$$\mathbf{T}_{\hat{G}}=\left\{\mathbf{f}=\left(f\left(x_1\right)+...+f\left(x_n\right)\right)\right\}+\boldsymbol{\varepsilon} \qquad (5)$$

where $\boldsymbol{\varepsilon}=\mathbf{T}_{\hat{G}}-\mathbf{f}$. Conceptually, as $n\to\infty$, $\mathrm{E}(\boldsymbol{\varepsilon})\cong 0$. This information will help us to define estimates for the function in (5). Then, kernel smoothing is can be used as a nonparametric approcah to get a proper estimate of the $f(.)$ 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 ith response observation. Then, a kernel smoother is defined as follows

$$\hat{T}_{i\hat{G}}=\sum_{j=1}^{n}w_{ij}t_j \qquad (6)$$

where t_j's are elements of the synthetic response variable $T_{i\hat{G}}$, and w_{ij}'s are known as kernel weights given by Nadaraya-Watson (1964). The specific weights for the kernel smoothing is expressed as

$$w_{ij}=K\left(\frac{x-x_j}{h}\right)\Big/\sum_{j=1}^{n}K\left(\frac{x-x_j}{h}\right)=K(u)/\sum K(u)$$

where h is a bandwidth parameter, and $\sum w_{ij}=1$. The function $K(u)$ determines the shape of the regression curves, while the parameter *h* determines their width.This approach is a called kernel smoothing because of a kernel function, *K*, to determine the weights. These kernel functions have the following properties: $K(u)\ge 0$ for all u, $K(u)=K(-u)$ and $\int K(u)du=1$. For example, Gaussian kernel function,

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

and other alternative kernel functions provide the properties of the kernel weight function, $K(u)$.

The kernel smoother (6) is also can be rewritten as in matrix form

$$\hat{\mathbf{T}}_{\hat{G}}=\mathbf{W}_h\mathbf{T}_{\hat{G}}=\hat{\mathbf{f}}_h \qquad (7)$$

where $\mathbf{W}_h=\left[w_{ij}\right]$ is a kernel smoother matrix based on parameter h. As in expressed before, the most important issue in this study is to select the bandwidth parameter. For this purpose, it is considered the most widely used three selection criteria, given in the next section

3 Bandwidth Selection Methods

Our task is to select an optimum value of the h.The optimum h is defined as the smoothing parameter which minimizes the average of the mean square errors (AMSE), given by

$$\mathrm{AMSE}(h)=\frac{1}{n}\left\|\left(\mathbf{I}-\mathbf{W}_h\right)\mathbf{T}\right\|^2+\frac{\sigma_\varepsilon^2}{n}tr\left(\mathbf{W}_h^2\right) \qquad (8)$$

where $\mathbf{W}_h$ is given in equation (7). The estimator of the error variance σ_ε^2 is described as follows:

$$\hat{\sigma}^2=n^{-1}\left\|\left(\mathbf{I}-\mathbf{W}_h\right)\mathbf{T}\right\|^2\Big/\left(n-p\right) \qquad (9)$$

where (*n-p*) is the degrees of freedom for residuals. Also note that $\hat{\sigma}^2$ is equal to mean square error (MSE) of the model. In this study, we used the MSE to measure the quality of estimated curves.

***Improved Akaike Criterion* (AICc):** This criterion is described by Hurvich et. al. (1998):

$$AIC_c = 1 + \log\left(\left\|(\mathbf{W}_h - \mathrm{I})\mathbf{T}\right\|^2 / n\right) + \left[\left(2tr(\mathbf{W}_h)+1\right)/n\right] - tr(\mathbf{W}_h) - 2 \quad (10)$$

***Generalized Cross-Validation Method* (GCV):** The GCV is defined by Craven and Wahba (1979):

$$\mathrm{GCV} = n^{-1}\left\|(\mathbf{I}-\mathbf{H}_h)\mathbf{T}\right\|^2 \Big/ \left[n^{-1}tr(\mathbf{I}-\mathbf{H}_h)\right]^2 \quad (11)$$

***Risk Estimation using Classical Pilots* (RECP):** The RECP score is expressed as

$$\mathrm{RECP} = \frac{1}{n}\left\{\left\|(\mathbf{W}_h - \mathbf{I})\hat{\mathbf{f}}_{h_p}\right\|^2 + \hat{\sigma}^2_{h_p} tr(\mathbf{W}_h\mathbf{W}_h')\right\} \quad (12)$$

where $\hat{\sigma}^2_{h_p}$ and $\hat{\mathbf{f}}_{h_p}$ are the appropriate *pilot estimates* for $\hat{\sigma}^2$ and $\hat{\mathbf{f}}$, respectively (Lee (2001) and Lee & Solo (1999)).

4 Simulation Experiment

This section reports a simulation experiment that evaluates the selection criteria given in Section 3. To see the performance of the small, medium and large samples of each criteria, we use three censoring levels (CLs), 15%, 35%, and 50% and three samples sizes with n = 50, 100, and 200. The number of replication was 1000 for each of the samples. The response observations are obtained by $T_i = f(x_i) + \varepsilon_i$, $1 \le i \le n$ where $\varepsilon_i \sim N(0, \sigma^2 = 1)$,

$$f(x_i) = 0.3\exp\left(-0.64(x_i - 0.25)^2\right) + 0.7\exp\left(-256(x_i - 0.75)^2\right)$$

and $x_i = (i - 0.5)/n$. Furthermore, we used the values of mean square error (MSE) to evaluate the quality of any curve estimate $(\hat{\mathbf{f}}_\lambda)$:

$$MSE = \frac{1}{1000}\sum_{i=1}^{n}\left\{f(x_i) - \hat{f}_h(x_i)\right\}^2, 1 \le i \le 1000 \quad (13)$$

The simulation experiment results are summarized in the following Table 1 and Figures 1-3.

Table 1: MSE values for nonparametric models

n	CLs	AICc	GCV	RECP
50	15	0.0248	0.0202	**0.0175**
	35	0.0362	0.0287	**0.0252**
	50	0.0366	0.0297	**0.0263**
100	15	0.0163	0.0148	**0.0133**
	35	0.0206	0.0188	**0.0162**
	50	0.0258	0.0237	**0.0199**
200	15	0.0143	0.0137	**0.0127**
	35	0.0178	0.0171	**0.0153**
	50	0.0182	0.0177	**0.0167**

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 RECP outperforms than the others for all censoring levels and samples size.

Boxplots for MSE values based on each criterion are illustrated in Figure 1. 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 GCV. Finally, G1, G2 and G3 indicate the MSE values for RECP. Also, the upper panel of Figure 1 has CL=15%, medium panel CL=35%, and bottom panel CL= 50%.

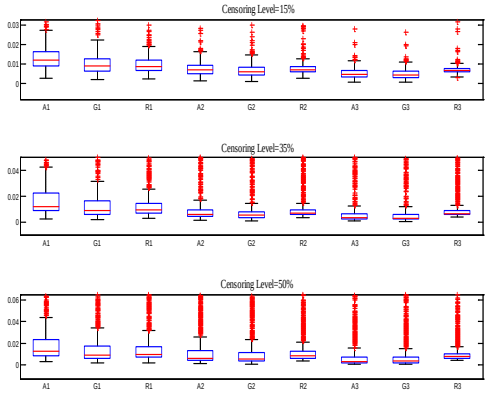

Figure 1: Boxplots of the MSE values for estimated nonparametric models

As can be seen 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

The left panel of the Figure 2 represents the curves estimated by AICc, GCV and RECP criteria for n=50 and CL=15%, while right panel shows the same curves but for CL= 50%.

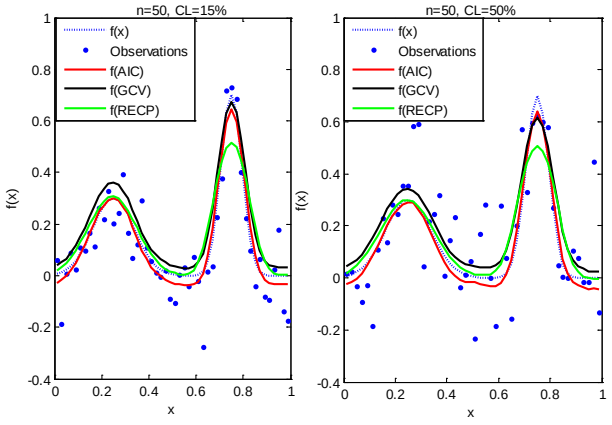

Figure 2: Real observations and the true function together with its smooth curves estimated by AICc, GCV and RECP

As for Figure 3, it is similar to Figure 2, the left panel represents the same curves for n=100 and CL=15%, while right panel shows the same curves for n=200 and CL= 50%.

Each panel compares the AICc, GCV and RECP fit to real functions. As can be seen from Figures 2-3, 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 are reasonable for censoring levels, CL=15% and 35%, when compared to the CL=50%.

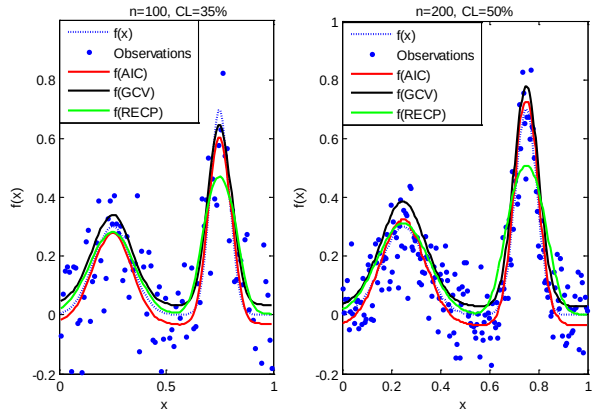

Figure 3: Similar to Figure 2, but for n=100, CL=35%, n=200 and CL=50%.

5 Conclusions

In this paper, we used kernel smoothing method to get the fits of an unknown regression function in nonparametric model with right censored data. Efficient computation of this method requires an optimum smoothing parameter. This parameter provided by means of AICc, GCV, and RECP criteria. Accordingly, we obtained three different estimators for the nonparametric regression function by using these criteria.We considered a simulated 1000 test observations to compare three different estimators for all sample sizes and censoring levels.

Consequently, the simulation results confirm that we can suggest the following main ideas:

- RECP criterion illustrates the better performance than the other criteria for all sample sizes and all censoring levels.
- Improved version of AIC and GCV criteria have similar performances in general, but GCV is the better than the improved AIC.
- Also as can be seen in Figures 2-3, although all estimated curves are close to real function, the curve estimated by RECP hardly distinguish from real function and this method gave the best values for all simulation study.

References:

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

[2] Koul, H., Susarla, V., Van Ryzin, J. (1981), Regression Analysis with Randomly Right-Censored Data, The Annals Of Statistics, 1276-1285.

[3] Nadaraya, E. A. (1964), On Estimating Regression, Theory Of Probability & Its Applications, Vol. 9(1), 141-142.

[4] Watson, G.S. (1964), Smooth Regression Analysis, Sankhya, Series A, Vol. 26, 359-372.

[5] Wand, M.P., Jones, M.C.(1984), Kernel Smoothing, Chapman & Hall.

[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] Hardle,W. Müller, M. (1997), Multivariate and Semiparametric Kernel Regression, Doctorate Dissertation, University Of Humboldt, Institute of Statistics and Econometrics, Berlin, Germany.

[9] Kaplan, E.L., Meier, P. (1958), Nonparametric Estimation From Incomplete Observations, Journal Of The American Statistical Association, Vol. 53(282), 457-481.

[10] Leurgans, S. (1987), Linear Models, Random Censoring and Sythetic Data, Biometrika, Vol. 74(2), 301-309.

[11] Zheng, Z.K. (1984), Regression Analysis with Censored Data, Ph.D Dissertation, University Of Colombia.

[12] 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.

[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] Lee, T.C.M. (2001), A stabilized Bandwidth Selection Method For Kernel Smoothing of Periodiagram, Signal Process, Vol. 81,419-430.

[15] Lee, T.C.M., Solo, V. (1999), Bandwidth Selection for Local Linear Regression: A Simulation Study, Computational Statistics and Data Analysis, Vol. 42, 139-148.

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