ESKİŞEHİR TECHNICAL UNIVERSITY JOURNAL OF SCIENCE AND TECHNOLOGY
A- APPLIED SCIENCES AND ENGINEERING

2020, Vol: 21, pp. 1-6, DOI: 10.18038/estubtda.817979

KERNEL SMOOTHING AS AN IMPUTATION TECHNIQUE FOR RIGHT-CENSORED DATA

Dursun AYDIN [1], **Ersin YILMAZ** [2,*]

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

ABSTRACT

Imputation of right-censored observations is an important problem in statistics and other applied sciences. Since right-censored data sets are common in medical studies and survival analysis, researchers should be careful about data quality. In this sense, imputation techniques are used to correctly estimate and complete censored data points. This study introduces the kernel smoothing method as an imputation method that takes into account the structure of the data and the individual effects of the accessible data points with kernel weights. The basic idea is to obtain a nonparametric model from the missing data set and consider sample predictions to estimate the censored ones. A simulation study is conducted to show the benefits of the method, and it is also compared with Ordinary Least Squares (OLS) based imputation, which is one of the widely used imputation methods and works similar to the proposed method.

Keywords: Kernel smoothing, right-censored data, Imputation technique, OLS imputation

1. INTRODUCTION

Imputation of right-censored observations is vital in statistical and other scientific fields. Since right-censored data sets are common in medical studies and survival analysis, researchers should be more meticulous about data quality. Imputation techniques are used to correctly estimate and complete the censored data values.

The purpose of this study is to introduce the kernel smoothing (KS) method as an imputation technique that takes into account the structure of the data and the individual effects of the data points accessible with the kernel weights. The basic idea is to derive a nonparametric model from the incomplete data set and make within-sample predictions to estimate the censored ones. A simulation study is conducted to show the benefits of the method, and it is also compared with OLS based imputation method, which is one of the widely used imputation methods and studies similar to the proposed method. The main contribution of the study is to obtain more data-driven data points using kernel weights than alternative OLS loading.

There are traditional methods to solve the problem of censorship in the literature. The most used ones are synthetic data transformation. Examples of these studies include [1-8]. In the context of the imputation, there are several notable studies such as [9-13]. The cited references examined different types of attribution techniques for both missing data and censorship data. The main difference of this article from the studies given is that it can be said that the introduced KS imputation method is more data-driven technique than the others. It allows using kernel weights.

The rest of the paper is designed as follows. Section 2 contains the methodology of the paper, Section 3 provides simulation experiments and results and finally, conclusions are given in section 4.

*Corresponding Author: yilmazersin13@hotmail.com
Received: 29.10.2020 Published: 27.11.2020

2. METHODOLOGY

We assume a sample of observations $\{x_i, y_i\}, i = 1,2,..,n$ and consider the nonparametric simple regression

$$y_i = g(x_i) + \varepsilon_i \tag{1}$$

where y_i's are the incompletely observed response values, x_i's are the values of covariate and $\varepsilon_i's$ are the error term that satisfies $E(\varepsilon|x) = 0$. It should be noted that the key idea is to estimate the unknown function $g(.)$ using response observations with missing values. This case state that ordinary nonparametric regression techniques cannot be used directly in the fitting procedure of the model (1). Hence, it is necessary to transform response observations (i.e., the values of response variable y).

[14-15] uses a kernel weights for modelling $g(.)$ as $\hat{g}(x) = \sum_{i=1}^{n} w_i(x) y_i = \mathbf{Wy}$, where $\mathbf{W}$ is a $n \times n$ dimensional weights matrix defined by

$$w_i(x) = \frac{K(x_i - x)}{h} / \sum_{i=1}^{n} \frac{K(x_i - x)}{h} = \frac{K(u)}{h} / \sum_{i=1}^{n} \frac{K(u)}{h} = \mathbf{W} \tag{2}$$

Here h is a bandwidth parameter, which is a nonnegative number controlling the size of the local neighborhood and $K(.)$ is a kernel function satisfying $K(u) = K(-u)$, $K(u) \geq 0$ and $\int K(u)du = 1$ for all $u \in R$. Note that we can define an estimate of $g(x)$ as $\hat{g}(x)$, where minimizes the weighted residual sum of squares (RSS)

$$RSS = \sum_{i=1}^{n} \big(y_i - g(x_i)\big)^2 w_i(x) \tag{3}$$

Thus, based on kernel smoothing, equation (3) is modified to take into account the presence of censored data by minimizing the following expression:

$$RSS = \sum_{i=1}^{n} \Big(z_{(i)} - g\big(x_{[i]}\big)\Big)^2 w_{[i]}(x) \tag{4}$$

where $z_{(1)}, \dots, z_{(n)}$ are the ordered values of the observed response variable $z_i = \min(y_i, c_i)$ in which y_i's are censored by censoring variable c_i's, $x_{[i]}$ is the value of the covariate associated with the ith ordered observation z_i, and $w_{[i]}$ is the kernel weight assigned to z_i .

Thus, equation (4) is minimized by

$$\hat{\mathbf{z}}_{KS} = \hat{\mathbf{g}}_{KS} = \mathbf{W}\mathbf{z}_G = \sum_{j=1}^{n} \big(w_{ij} z_{jG}\big) \mathrm{I}[\delta_i = 1] \tag{5}$$

where $\hat{\mathbf{z}}_{KS}$ is a $(n \times 1)$ vector of fitted values obtained by kernel smoothing, $\delta_i = \mathrm{I}(y_i \leq c_i)$ is the value of the non-censored indicator variable ($\delta_i = 1, if\ y_i \leq c_i$ otherwise 0) associated with the ith ordered observation z_i. Note also that z_i is censored then $\delta_i = 0$ and if it is not censored, $\delta_i = 1$. In addition, KS estimator defined in (5) of the vector $\mathbf{g}$ based on synthetic responses can be written in matrix and vector form

$$\hat{\mathbf{z}}_{KS} = \hat{\mathbf{g}}_{KS} = \mathbf{W}\mathbf{z}_{\mathrm{G}}\, \mathrm{I}[\boldsymbol{\delta} = 1] \tag{6}$$

It appears from (6) that the fitted values obtained for each uncensored observation are replaced by incomplete observations. Procedure of KS imputation is given in Algorithm 1.

Algorithm 1: ***KS imputation for right-censored data***

Input: Right-censored dataset z_i
Censoring indicator δ_i
Values of predictor variable x_i
Output: Imputed dataset $\hat{\mathbf{y}}^{ks} = (\hat{y}_1^{ks}, \dots, \hat{y}_n^{ks})^T$
1 **begin**
2 **for** (*i*=1 to *n*) **do**
3 **if** $\delta_i = 1$ **do** (if observation is uncensored)
4 obtain z_i^{ks} with $z_i^{ks} = z_i$
5 obtain x_i^{ks} with $x_i^{ks} = x_i$
6 **end** (for the loop in step 2)
7 Estimate the fitted values $\hat{\mathbf{z}}^{ks}$ using kernel smoothing method by $\hat{\mathbf{z}}_{KS} = \hat{\mathbf{g}}_{KS} = \mathbf{W}\mathbf{z}_G\ \mathbf{I}[\boldsymbol{\delta} = 1]$
8 **for** (*i*=1 to *n*) **do**
9 **if** $\delta_i = 0$ **do**
10 Estimate right-censored observation for x_i value and obtain $\hat{y}_i^{ks}$
11 Replace the estimated value $(\hat{y}_i^{ks})$ with censored one (z_i)
12 **end** (for the loop in step 8)
13 Return $\hat{\mathbf{y}}^{ks} = (\hat{y}_1^{ks}, \dots, \hat{y}_n^{ks})^T$
14 **end**

Linear regression is a classical method of modeling response variables with a linear equation based on the estimated regression coefficients. Let *m* be the number of uncensored data points and z_i^{ols} be the value(s) of uncensored observations. In this case, the general form of linear regression model as follows

$$z_i^{ols} = \beta_0 + \beta_1 x_1^{ols} + \cdots + \beta_p x_p^{ols} + \varepsilon_i^{ols}, i = 1, \dots, m \tag{7}$$

The minimization problem to give optimal estimates of the regression coefficients β in (7) can be written as

$$OLS(\beta^{ols}) = \sum_i^n \left[z_i^{ols} - \left(x_i^{ols}\right)^T \beta^{ols} \right]^2 \tag{8}$$

In general, the solution of equation above is provided by gradient descent, given by

$$\widehat{\boldsymbol{\beta}}^{ols} = \left[\left(\mathbf{x}^{ols}\right)^T \mathbf{x}^{ols} \right]^{-1} \left(\mathbf{x}^{ols}\right)^T \mathbf{z}^{ols} \tag{9}$$

An algorithm for the generation of a predictive model is given by

Algorithm 2: ***OLS imputation for right-censored data***

Input: Right-censored dataset z_i
Censoring indicator δ_i
Values of predictor variable x_i
Output: Imputed dataset $\hat{\mathbf{y}}^{ols} = (\hat{y}_1^{ols}, \dots, \hat{y}_n^{ols})^T$
1 **begin**
2 **for** (*i*=1 to *n*) **do**
3 **if** $\delta_i = 1$ **do** (if observation is uncensored)
4 obtain z_i^{ols} with $z_i^{ols} = z_i$
5 obtain x_i^{ols} with $x_i^{ols} = x_i$
6 **end** (for the loop in step 2)
7 Estimate the vector of the regression coefficients $\boldsymbol{\beta}^{ols}$ with Eq. (9)
8 **for** (*i*=1 to *n*) **do**
9 **if** $\delta_i = 0$ **do**
10 Estimate right-censored observation for x_i value with using $\widehat{\boldsymbol{\beta}}^{ols}$ and obtain $\hat{y}_i^{ols}$
11 Replace the estimated value $\left(\hat{y}_i^{ols}\right)$ with censored one (z_i)
12 **end** (for the loop in step 8)
13 Return $\hat{\mathbf{y}}^{ols} = (\hat{y}_1^{ols}, \dots, \hat{y}_n^{ols})^T$
14 **end**

2.1. Performance Measures

Three different criteria are considered: averaged-bias (AvB) for the imputed values, root mean squared error (RMSE), and inaccuracy (IA) measure. These are defined, respectively, as

$$AvB = \frac{1}{n_{cens}} \sum_{i=1}^{n_{cens}} \left| y_i - y_i^{imp} \right| \quad RMSE = \sqrt{\frac{1}{n_{cens}} \sum_{i=1}^{n_{cens}} \left(y_i - y_i^{imp} \right)^2}$$

$$IA = \frac{1}{n_{cens}} \sum_{i=1}^{n_{cens}} |y^* - y_i^{imp}| / y_i$$

Note that IA measures the inaccuracy of imputation defined as the mean of the proportional difference between real and imputed values.

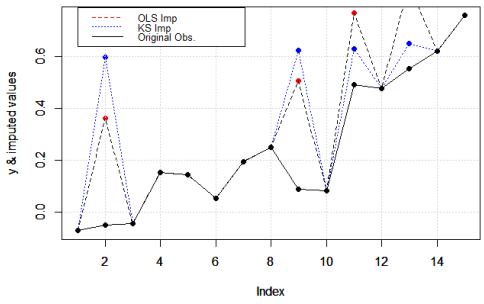
3. SIMULATION STUDY

Simulation study is designed as follows:

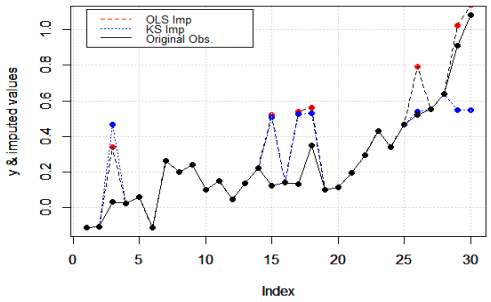
- Covariate x_i's are obtained from the uniform distribution on $U[0,1]$.
- To introduce right censoring, we generate the censoring variable c from the normal distribution on $N(\mu_y, \sigma_y)$.
- The censoring indicator $\delta = I(y \leq c)$ is generated from Bernoulli distribution with proportions at censoring rates such as 5%, 25%, and 50%.
- For each censoring rates in simulation, we construct 1000 simulated random samples of size n=15, 30, and 75.

Results are presented in Figure 1 and Table 1.

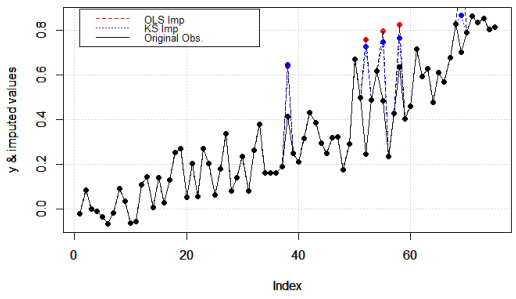
(a) n=15, CL=5%



(b) n=30, CL=25%



(c) n=75, CL=5%



(d) n=75, CL=45%

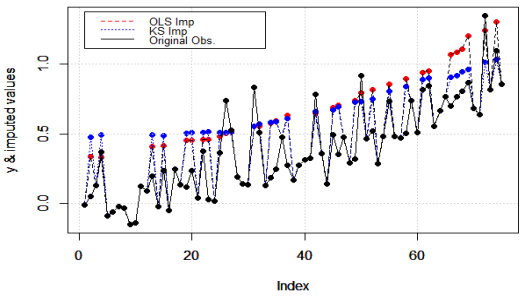


Figure 1. Comparative figures for imputation methods with original observations (black), KS imputation (blue) and OLS imputation (red).

Figure 1 involves output figures from different combinations. As can be seen results are closer to each other but in detail, it can be seen that KS has better performance than OLS imputation almost for all simulation experiments. These results are proved by Table 1 given in the next slide.

Table 1. Performance scores for all configurations

		The imputed variables (y_i^{imp})					
n	*C.L*	IA_{KS}	IA_{OLS}	AvB_{KS}	AvB_{OLS}	$RMSE_{KS}$	$RMSE_{OLS}$
15	5%	0.0935	**0.0663**	**0.0200**	0.0202	**0.0071**	0.0073
	25%	**0.0597**	0.0695	**0.0213**	0.0217	**0.0116**	0.0118
	45%	**0.0569**	0.0588	0.0221	**0.0200**	**0.0016**	0.0017
30	5%	**0.0467**	0.0491	0.0196	0.0196	**0.0066**	0.0068
	25%	0.0753	**0.0669**	0.0205	**0.0201**	0.0118	**0.0115**
	45%	0.0409	**0.0364**	0.0200	0.0200	**0.0147**	0.0157
75	5%	0.1903	**0.1833**	**0.0197**	0.0201	0.0078	**0.0073**
	25%	**0.0523**	0.0689	0.0208	**0.0201**	**0.0105**	0.0115
	45%	**0.0271**	0.0330	**0.0157**	0.0198	**0.0157**	0.0159

Table 1 involves scores of IA, AvB and $RMSE$ values for both imputation techniques. The best scores are indicated with bold color. As can be seen in Table 1, KS imputation gives more bold scores than OLS. It can be explained with its Sensitivity to data points and its nonparametric nature.

4. CONCLUSIONS

In this paper, we introduce a new imputation technique for the right-censored data based on kernel smoothing technique. An algorithm is developed to apply the method to data and to measure its performance, it is compared by its old alternative OLS imputation. To achieve proper comparison a simulation study is realized for different setup. The outcomes of simulation study prove that KS imputation gives better results than OLS imputation. The ability of KS imputation on completing censored data points can be explained by kernel weights that allow the data driven estimation. In addition, Table 1 and Figure 1 support that KS imputation provides some benefits to right-censored literature.

REFERENCES

[1] Koul H, Susarla V, Van Ryzin J. Regression Analysis with Randomly Right-Censored Data. The Annals of Statistics, 1981; 9(6), 1276-1285.

[2] Leurgans S. Linear models, random censoring and synthetic data. Biometrika, 1987; 74, 301-309.

[3] Buckley J, James, I. Linear Regression with Censored Data. Biometrika, 1979; 66(3), 429-436.

[4] Aydın D, Yılmaz, E. Modified Spline Regression Based On Randomly Right-Censored Data: A Comparison Study, Communications in Statistics-Simulation and Computation, 2017; doi: 10.1080/03610918.2017.1353615.

[5] Kaplan EL and Meier P. Nonparametric Estimation from Incomplete Observations, Journal of the American Statistical Association, 1958; 53(282), 457-481.

[6] Miller RG. Least squares regression with censored data. Biometrika, 1976; 63, 449-64.

[7] Stute W. Consistent Estimation under Random Censorship When Covariables are Present. Journal of Multivariate Analysis, 1993; 45, 89-103.

[8] Aydın D and Yılmaz E. Right-censored nonparametric regression: A comparative simulation study, TEM Journal, 2016; 5(4), 446-450.

[9] Hasler C and Craiu RV. Nonparametric imputation method for nonresponse in surveys, Statistical Methods & Applications, 2019; 29, 25-48.

[10] Musil CM, Warner CB, Yobas PK and Jones SL. A comparison of imputation techniques for handling missing data, Western Journal of Nursing Research, 2002; 24(7), 815-829.

[11] Chen L and Sun J. A multiple imputation approach to the analysis of interval-censored failure time data with the additive hazards model, Comput. Stat. Data. Anal., 2010; 54(4), 1109-1116.

[12] Ahmed SE, Aydın D and Yılmaz E. Nonparametric regression estimates based on imputation techniques for right-censored data, ICMSEM 2019: Proceedings of the Thirteenth International Conference on Management Science and Engineering Management, 2020; 109-120.

[13] Wei GCG and Tanner MA. Applications of multiple imputation to the analysis of censored regression data, Biometrics, 1991; 47(4), 1297-1309.

[14] Nadaraya EA. On Estimating Regression. Theory of Probability and Its Applications, 1964; 10, 186-190.

[15] Watson G S. Smooth regression analysis. Sankhya A 1964; 26, 359-72.