Türkiye Klinikleri Biyoistatistik Dergisi
Turkiye Klinikleri Journal of Biostatistics Turkiye Klinikleri J Biostat. 2022;14(2):90-102

ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

DOI: 10.5336/biostatic.2022-89354

Modelling Right-Censored Data with Partially Linear Model and Feed Forward Neural Networks: A Methodological Study

Sağdan-Sansürlü Verilerin Kısmi Doğrusal Model ve İleri Beslemeli Yapay Sinir Ağları ile Modellenmesi: Metodolojik Bir Çalışma

Çağatay BAL[a], Ersin YILMAZ[a]

[a]Department of Statistics, Muğla Sıtkı Koçman University Faculty of Art, Muğla, Türkiye

ABSTRACT Objective: Modeling right-censored data becomes a challenging task in survival analysis, due to having an incomplete data structure. When the response variable is right-censored, classical estimation methods cannot be used directly. Therefore, the censorship problem should be solved before the modeling process. The purpose of this study is to solve the censorship problem with synthetic data transformation and to make a comparison between the partial linear model (PLM) and feed forward neural networks (FFNN), two popular modeling procedures in recent years, in terms of model residuals. Thus, it is to study the behavior of methods. **Material and Methods:** This paper aims to estimate the effects of explanatory variables on a right-censored response variable whose distribution is unknown by two different methods, PLM and FFNN based on the spline smoothing method. The spline smoothing method is a mathematical approximation method used in PLM estimation. FFNN is a machine learning method that has become very popular recently and produces satisfactory models. To overcome the censorship problem, the right-censored response variable for the two mentioned methods has been replaced with synthetic data. Synthetic data transformation is a widely used censorship resolution method that allows censorship presence to be included in the prediction process. **Results:** To achieve the aim of the study, both simulation and real data studies were carried out and the results were presented. **Conclusion:** Kidney weakness data is used as an example of real data. When the results are examined, it is seen that FFNN is superior to PLM in both numerical studies.

Keywords: Feed forward neural networks; smoothing spline; partially linear models; right-censored data

ÖZET Amaç: Eksik bir veri yapısına sahip olması nedeniyle sağdan-sansürlenmiş verilerin modellenmesi, sağkalım analizinde zor bir işlemdir. Yanıt değişkeni sağdan sansürlendiğinde, klasik tahmin yöntemleri doğrudan kullanılamaz. Bu nedenle modelleme sürecinden önce sansür problemi çözülmelidir. Bu çalışmanın amacı, sansür problemini sentetik veri dönüşüm ile çözerek literatürde son yıllarda popüler olarak kullanılan 2 modelleme prosedürü olan kısmi doğrusal model [partial linear model (PLM)] ve ileri beslemeli sinir ağları [feed forward neural networks (FFNN)] arasında model artıkları açısından bir karşılaştırma yapmak ve böylece yöntemlerin davranışlarını incelemektir. **Gereç ve Yöntemler:** Bu makale, açıklayıcı değişkenlerin, dağılımı bilinmeyen sağdan sansürlü bir yanıt değişkeni üzerindeki etkilerini, splayn düzleştirme yöntemine dayalı PLM ve FFNN olmak üzere 2 farklı yöntemle tahmin etmeyi amaçlar. Splayn düzleştirme yöntemi, PLM tahmininde kullanılan matematiksel yaklaştırma yöntemidir. FFNN ise son zamanlarda oldukça popülerleşen ve tatmin edici modeller üreten bir makine öğrenmesi yöntemidir. Sansür sorununun üstesinden gelmek için bahsedilen 2 yöntem için sağdan sansürlü yanıt değişkeni sentetik verilerle değiştirilmiştir. Sentetik veri dönüşümü, sansür varlığını tahmin sürecine dâhil edilmesini sağlayan yaygın kullanılan bir sansür çözüm yöntemidir. **Bulgular:** Çalışmanın amacına ulaşmak için hem simülasyon hem de gerçek veri çalışmaları yapılmış ve sonuçlar sunulmuştur. **Sonuç:** Gerçek veri örneği olarak böbrek zayıflığı verisi kullanılmıştır. Sonuçlar incelendiğinde, her iki sayısal çalışmada da FFNN'nin PLM'ye üstünlük sağladığı açıkça görülmektedir.

Anahtar kelimeler: İleri beslemeli sinir ağları; splayn düzleştirme; kısmi doğrusal modeller; sağdan-sansürlü veri

Censored data arises in different fields such as medical, biology, public health, epidemiology, engineering, economics and, so on. Observations obtained from mentioned fields are usually incomplete, especially in medical studies. For instance, some patients may still be alive, disease-free, or die at the completion of a medical study. Thus, analysis of censored survival data becomes a common problem in the literature.

Correspondence: Ersin YILMAZ
Department of Statistics, Muğla Sıtkı Koçman University Faculty of Art, Muğla, Türkiye
E-mail: yilmazersin13@hotmail.com

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 01 Mar 2022 ***Received in revised form:*** 20 Apr 2022 ***Accepted:*** 28 Apr 2022 ***Available online:*** 10 May 2022

In this paper, estimation of right-censored data is carried out by both popular methods which are feed forward neural networks (FFNN) and semi-parametric regression model based on smoothing spline (SS), and their performances are compared.

In the literature, there are number of studies about semi-parametric regression based on SSs and estimation of right-censored data. In this context, estimation of semi-parametric regression models under the assumption of the response variable is observed completely and studied by several researchers.[1,2] There are some important studies about regression with right-censored data.[3,4] For example, the asymptotic properties of the right-censored semi-parametric regression estimators are discussed.[5] Recently, the censored partial regression with right-censored data and a new bootstrap estimation procedure is proposed based on weighted least squares.[6] Also, three smoothing methods for estimating the right-censored semi-parametric regression models are compared.[7]

Statistical methods such as piecewise parametric and semi-parametric regression models have been used successfully in the survival data analysis in clinical medicine. However, the evolution of existing methods and finding new approaches inevitably motivate the researchers to use them for better solutions. As one of the powerful nonlinear tools, neural networks have wide application areas due to their adaptation capabilities. The Neural networks are used for modeling survival data in several studies based on different concepts such as Bayesian networks and partial logistic models.[8] Also, some studies about survival data on the Bayesian analysis framework are realized.[9]

In addition, Biganzoli et al. proposed a framework for FFNN on censored survival data.[10] Azadeh et al. used neural networks and compared with fuzzy regression, exponential, and Weibull distributions methods for right-censored survival data analysis.[11] Kalderstam et al. used the genetic algorithm as an optimization algorithm of neural networks.[12]

The paper is organized as follows. In Section 2, the partially linear model and SS method is introduced. Also, FFNN is expressed in this section. A simulation study and real data work are illustrated in Sections 3 and 4. Finally, the discussion and conclusions are presented in Section 5 and 6 respectively.

MATERIAL AND METHODS

PARTIALLY LINEAR MODEL BASED ON SS WITH RIGHT-CENSORED DATA

Let semi-parametric regression model be as follows

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + f(t_i) + \varepsilon_i, \quad i = 1, \dots, n \qquad (2.1)$$

where the y_i's represent the values of the response variable, $\mathbf{x}_i = \left(x_{i1}, \dots, x_{ip}\right)^T$ are predictor variables that are completely observed with $p \leq n$, $\boldsymbol{\beta} = \left(\beta_1, \dots, \beta_p\right)^{\mathrm{T}}$ is the vector of the regression coefficient. Note that here, $\mathbf{x}_i\boldsymbol{\beta}$ is the parametric component of the model (2.1). Values of t_i represent the nonparametric explanatory variable, $f(.)$ is the unknown regression function to be estimated and finally, ε_i's are the random errors that have zero mean and constant variance.

Vector and matrix form of Model (2.1) can be rewritten as follows

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{f} + \boldsymbol{\varepsilon} \qquad (2.2)$$

where $\mathbf{Y} = (y_1, \dots, y_n)'$, $\mathbf{X}_i = (x_{i1}, \dots, x_{in})'$, $\mathbf{f} = (f(t_i), \dots, f(t_n))'$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)'$. Note that in this study only response variable $\mathbf{Y}$ is censored from the right by a random censoring variable c_i and it is assumed that all the explanatory variables are fully observed.

The main interest of this paper is estimating the parameter vector $\boldsymbol{\beta}$ and smooth function $\mathbf{f}$ when y_i's are observed incompletely also without the knowledge of the distributions of y and c variables and comparing the obtained results with Artificial Neural Network (ANN) estimates. In this context, instead of observing triplets (x_i, t_i, y_i), quartets $(x_i, t_i, z_i, \delta_i)$ are observed where

$$z_i = min\,(y_i, c_i) \text{ and } \delta_i = \begin{cases} 1;\, if\ y_i \leq c_i\ (observed) \\ 0;\, if\ y_i > c_i\ (censored) \end{cases} \quad (2.3)$$

It is assumed that in equation (2.3) y_i's and c_i's have identically and independently distributed (*i.i.d.*) variables with *F* and *G* distributions respectively.

There are two important assumptions that have to be ensured for the accuracy of the estimation which is expressed by Stute with details; (A1) y_i's and c_i's have to be independent and (A2) $P(y_i \leq c_i | x_i, t_i, y_i) = P(y_i \leq c_i | y_i)$.[13] The first assumption is the standard independence condition and the second assumption implies that given a lifetime, covariates cannot provide any further information as to whether the observation is censored or not.

Under these conditions, there are two methods to estimate the regression model that are weighted least squares with Kaplan and Meier weights and synthetic data transformation.[14] In this paper, synthetic data transformation proposed by Koul et al. is used who discussed that if the distribution of censoring times c_i's is continuous and known, an adjustment can be made on lifetimes z_i to obtain synthetic observations as follows:

$$z_{iG} = \frac{\delta_i z_i}{1 - G(z_i)}, \qquad i = 1, \dots, n \quad (2.4)$$

In practice, the distribution of the censoring times *G* is generally unknown.[3] To overcome this problem, *G* is replaced by its Kaplan and Meier estimator $\hat{G}$ which is given in equation (2.5),

$$1 - \hat{G}(z) = \prod_{j=1}^{n} \left(\frac{n-j}{n-j+1}\right)^{I[z_{(j)} \leq z, \delta_{(j)} = 0]}, (z \geq 0) \quad (2.5)$$

where $z_{(1)} \leq z_{(2)} \leq \cdots \leq z_{(n)}$ are the ordered values of the response variable and $\delta_{(i)}$'s are the corresponding censoring indicator associated with ordered z_i's.[14] In this case, equation (2.4) can be rewritten as follows:

$$\mathbf{Z}_{\hat{G}} = (z_{1\hat{G}}, \dots, z_{n\hat{G}})' = \frac{\delta_i z_i}{1 - \hat{G}(z_i)} = z_{i\hat{G}}, i = 1, \dots, n \quad (2.6)$$

In this paper, $z_{i\hat{G}}$ is called as synthetic observations. In this context, mentioned two assumptions provide that $E[z_{iG} | x_i, t_i] = E[z_i | x_i, t_i] = \mathbf{X\beta} + \mathbf{f}$.

Proof: $E[z_{iG} | x_i, t_i] = E[z_i | x_i, t_i]$

Derivation of the

$$\begin{aligned} E[z_{iG} | x_i, t_i] &= E\left[\frac{\delta_i z_i}{1 - G(z_i)} | x_i, t_i\right] = E\left[\frac{I(y_i \leq c_i)\, min(y_i, c_i)}{1 - G[min(y_i, c_i)]} | x_i, y_i\right] \\ &= E\left[I(y_i \leq c_i) \frac{y_i}{1 - G(z_i)} | x_i, t_i\right] = E\left[E\left[\frac{y_i}{1 - G(z_i)} I(y_i \leq c_i) | y_i, x_i, t_i\right] | x_i, t_i\right] \\ &= E\left[\frac{y_i}{1 - G(z_i)} \big(1 - G(z_i)\big) | x_i, t_i\right] = E(y_i | x_i, t_i) \end{aligned}$$

Hence, the proof is realized.

After synthetic data transformation, models (2.1-2.2) are given by

$$z_{i\hat{G}} = \mathbf{x}_i \boldsymbol{\beta} + f(t_i) + \varepsilon_{i\hat{G}}, \quad i = 1, \dots, n \quad (2.7)$$

and

$$\mathbf{Z}_{\hat{\mathbf{G}}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{f} + \boldsymbol{\varepsilon}_{\hat{\mathbf{G}}} \qquad (2.8)$$

where $\varepsilon_{i\hat{G}}$'s are identical but not independent error terms. Note that when $n \to \infty$, $E(\varepsilon_{i\hat{G}}) \cong 0$ which is necessary information for the accuracy of estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{f}}$.

***Proof**: when $n \to \infty$, $E(\varepsilon_{i\hat{G}}) \cong 0$*

For known cumulative distribution G, $E[z_{iG}|x_i, t_i] = E[z_i|x_i, t_i]$ can be implied. From here, expected values of error terms can be written,

$$E(\varepsilon_{iG}) = E[z_{iG} - E(z_{iG}|x_i, t_i)]$$
$$= E(z_{iG}) - E[E(z_{iG}|x_i, t_i)] = 0$$

Generally, distribution G is unknown, and its Kaplan and Meier estimate is shown in equation (2.5). Koul et al. discussed the approximation of $\hat{G}$ to G as $n \to \infty$.[3,14] From that, it can be written as

$$E(\varepsilon_{i\hat{G}}) = E(z_{i\hat{G}}) - E[E(z_{i\hat{G}}|x_i, t_i)] \cong 0$$

as claimed.

SS

This method works based on a penalized least square. In this context, let $k_1 < k_2 < \cdots < k_q$ be the ordered and unique values of the $t_1, t_2, \ldots, t_n$. Here, q is the number of unrepeated values of variable *t.* The connection between t_i's and k_i's provided by $n\ x\ q$ incidence matrix (**N**). Elements of the incidence matrix are formed as $N_{ij} = 1\ if\ t_i = k_j$ and $N_{ij} = 0\ if\ t_i \neq k_j$.

Estimators can be obtained by minimizing the equation given below;

$$L(\boldsymbol{\beta}; \mathbf{f}) = (\mathbf{Z}_{\hat{G}} - \mathbf{X}\boldsymbol{\beta} - \mathbf{N}\mathbf{f})'(\mathbf{Z}_{\hat{G}} - \mathbf{X}\boldsymbol{\beta} - \mathbf{N}\mathbf{f}) + \lambda \int_a^b f''(t)^2 dt \qquad (2.9)$$

where $\lambda > 0$ is the smoothing parameter that controls the penalty term and smoothness of the regression function. In equation (2.9), the first term represents the goodness of fit to data and the second term $\lambda \int_a^b f''(t)^2 dt$ is called as a penalty term.

From the properties of the SS, the penalty term can be written as $\int_a^b f''(t)^2 dt = \mathbf{g}'\mathbf{K}\mathbf{g}$.[1] Here, **K** is the symmetric and positive definite $q\ x\ q$ matrix and defined as $\lambda\mathbf{K} = \mathbf{S}_\lambda^{-1} - \mathbf{I}$ where **I** is a unit matrix and $\mathbf{S}_\lambda$ is a smoother matrix that determines the shape of the nonparametric function with smoothing parameter λ. Note that, **K** can be calculated as $\mathbf{K} = \mathbf{Q}'\mathbf{R}^{-1}\mathbf{Q}$. Here, **Q** and **R** are tri-diagonal $(q-2)\ x\ q$ and $(q-2)\ x\ (q-2)$ dimensional matrices respectively. Elements of **Q** and **R** are indicated below;

$$Q_{ii} = \frac{1}{h_i}, \quad Q_{i,(i+1)} = -\left[\frac{1}{h_i} + \left(\frac{1}{h_{(i+1)}}\right)\right], \quad Q_{i,(i+2)} = \frac{1}{h_{(i+1)}}$$
$$R_{(i-1),i} = R_{i,(i-1)} = \frac{h_i}{6}, \qquad R_{ii} = \frac{\left(h_i + h_{(i+1)}\right)}{3}$$

where $h_i = k_{(i+1)} - k_i,\ i = 1,2, \ldots, (q-1)$. From that solution of the minimization problem (2.9) ensures the estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{f}}$ that are calculated as follows:

$$\hat{\boldsymbol{\beta}} = [\mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{X}]^{-1}\mathbf{X}'\left((\mathbf{I} - \mathbf{S}_\lambda)\mathbf{Z}_{\hat{G}}\right) \qquad (2.10a)$$

and

$$\hat{\mathbf{f}} = (\mathbf{N}'\mathbf{N} + \lambda\mathbf{K})^{-1}\mathbf{N}'\left(\mathbf{Z}_{\hat{G}} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) \qquad (2.10b)$$

In here, $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{f}}$ are called as modified estimators of $\boldsymbol{\beta}$ and $\mathbf{f}$ for semi-parametric regression model under right-censored data.

Properties of Estimators

Expansion of the $\widehat{\boldsymbol{\beta}}$ could be written as follows:

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\widetilde{\mathbf{X}}\right)^{-1}\mathbf{X}'\widetilde{\mathbf{Z}}_{\hat{G}} = \boldsymbol{\beta} + \left(\mathbf{X}'\widetilde{\mathbf{X}}\right)^{-1}\mathbf{X}'\widetilde{\mathbf{f}} + \left(\mathbf{X}'\widetilde{\mathbf{X}}\right)^{-1}\mathbf{X}'(\mathbf{I}-\mathbf{S}_{\lambda})\boldsymbol{\varepsilon}_{\hat{G}}$$

where $\widetilde{\mathbf{X}} = (\mathbf{I}-\mathbf{S}_{\lambda})\mathbf{X}$, $\widetilde{\mathbf{Z}}_{\hat{G}} = (\mathbf{I}-\mathbf{S}_{\lambda})\mathbf{Z}_{\hat{G}}$ and $\widetilde{\mathbf{g}} = (\mathbf{I}-\mathbf{S}_{\lambda})\mathbf{f}$. From those calculations of the bias and variance-covariance matrix can be given as

$$Bias(\widehat{\boldsymbol{\beta}}) = E(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\beta} = \left(\mathbf{X}'\widetilde{\mathbf{X}}\right)^{-1}\mathbf{X}'\widetilde{\mathbf{f}} \quad (2.11)$$

$$Var(\widehat{\boldsymbol{\beta}}) = \sigma^2\left(\mathbf{X}'\widetilde{\mathbf{X}}\right)^{-1}\mathbf{X}'(\mathbf{I}-\mathbf{S}_{\lambda})^2\mathbf{X}\left(\mathbf{X}'\widetilde{\mathbf{X}}\right)^{-1} \quad (2.12)$$

σ^2 is generally unknown in the real world. To overcome this difficulty, the estimation of σ^2 is needed that obtained via error terms. In this case, the residual sum of squares (RSS).

$$RSS = \sum_{i=1}^{n}\varepsilon_{\hat{G}}^2 = \left(\mathbf{Z}_{\hat{G}} - \widehat{\mathbf{Z}}_{\hat{G}}\right)'\left(\mathbf{Z}_{\hat{G}} - \widehat{\mathbf{Z}}_{\hat{G}}\right) \quad (2.13)$$

Here, replacing $\widehat{\mathbf{Z}}_{\hat{G}} = \mathbf{H}_{\lambda}\mathbf{Z}_{\hat{G}}$, equation (2.12) is updated as below

$$RSS = (\mathbf{Z}_{\hat{G}} - \mathbf{H}_{\lambda}\mathbf{Z}_{\hat{G}})'(\mathbf{Z}_{\hat{G}} - \mathbf{H}_{\lambda}\mathbf{Z}_{\hat{G}}) = \|(\mathbf{I}-\mathbf{H}_{\lambda})\mathbf{Z}_{\hat{G}}\|^2 \quad (2.14)$$

where $\mathbf{H}_{\lambda}$ is the hat matrix for the semi-parametric regression model under right-censored data. From that estimator of σ^2 can be obtained as follows:

$$\hat{\sigma}^2 = \frac{RSS}{tr(\mathbf{I}-\mathbf{H}_{\lambda})^2} = \frac{RSS}{(n-p)} \quad (2.15)$$

where $tr(\mathbf{I}-\mathbf{H}_{\lambda})^2$ represents the degree of freedom for the model and p is the number of the active parameter in the model.

Feed Forward Neural Networks

FFNN is a type of multi-layer perceptron which has an interconnection between all neurons in a network and unlike different types of multi-layer perceptrons, FFNN has no loops or circles within the architecture and signal flows through input layer to output layer in one direction. FFNN architecture presentation can be seen in Figure 1.

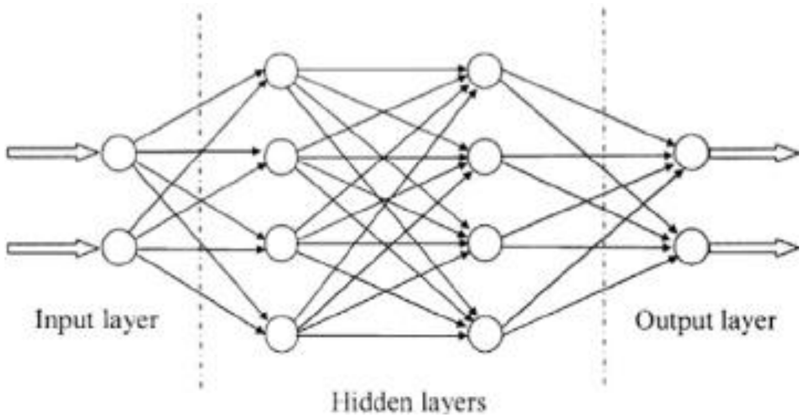

FIGURE 1: Feed forward neural networks architecture presentation.

Since the late ‘80s after the discovery of the backpropagation algorithm, many studies proved that the neural networks as a type of non-linear modeling tool is a strong alternative to other conventional statistical methods.[15] Neural networks models can give very favorable solutions, especially in the case of the presence of complex, noisy, irrelevant, or partial information.[16,17]

In this study, FFNN architecture with two hidden layers has been set for the main component of the neural network. The reason behind the idea is the inconsistent results of one hidden layered network with multiple trials of modeling. The range of the errors of the outputs among trials is considerably high and therefore one hidden layered network cannot promise a reliable modeling solutions for this complex structure of censored data.

The general framework of backpropagation can be described as follows:

$$E_p = \sum_k (d_k - x_k)^2 \quad (2.16)$$

Error function E_p defines error between target value (d_k) and output value (x_k) of k^{th} neuron in the network.

$$\overline{\in}_i = \frac{\partial^+ E}{\partial \bar{x}_i} \quad (2.17)$$

Gradient vector of errors $\overline{\in}_i$ will be obtained for i^{th} neuron in the network.

$$\overline{\in}_i = \begin{cases} -2(d_i - x_i)\frac{\partial x_i}{\partial \bar{x}_i} = -2(d_i - x_i)x_i(1-x_i), if\ i \text{ ith neuron is output neuron} \\ \frac{\partial x_i}{\partial \bar{x}_i} = \sum_{j,i<j} \frac{\partial^+ E}{\partial \bar{x}_j}\frac{\partial \bar{x}_j}{\partial x_i} = x_i(1-x_i)\sum_{j,i<j} \overline{\in}_j\, w_{ij}\ , \quad otherwise \end{cases} \quad (2.18)$$

In equation (2.18), w_{ij} is the weights between i^{th} and j^{th} neuron in network. If this value equals zero then that means, there is no connection between i^{th} and j^{th} neuron in the network.

$$\Delta w_{ki} = -\eta \frac{\partial^+ E_p}{\partial w_{ki}} = -\eta \frac{\partial^+ E_p}{\partial \bar{x}_i}\frac{\partial \bar{x}_i}{\partial w_{ki}} = -\eta\ \overline{\in}_j\ x_k \quad (2.19)$$

In equation (2.18), η is described as the learning ratio which affects the convergence speed and the stability of the weights in the learning process. Bias can be updated in the same way as equation (2.19).

All weights are obtained after each iteration in the training process.

$$\Delta w_{ki} = -\eta \frac{\partial^+ E}{\partial w_{ki}} = -\eta \sum_p \frac{\partial^+ E_p}{\partial w_{ki}} \quad (2.20)$$

$$\Delta w = -\eta \frac{\partial^+ E}{\partial w} = -\eta \nabla_w E \quad (2.21)$$

In equations (2.20-2.21), weights updating through error gradients $E = \sum_p E_p$ are described and the gradients will be calculated throughout the data set.

In this study, the Levenberg-Marquadt backpropagation algorithm has been chosen for its fast convergence speed as a learning algorithm of FFNN.

Activation functions are an important component of neural networks that allows non-linear mapping between input and target values in the dataset.

Tangent and logistic sigmoidal functions are the most common activation functions for hidden layer because of their `S` shapes that allow flexible range for mapping and also non-negative derivatives between ranges of the functions as [-1,1] for tangent-sigmoid and [0,1] for logistic-sigmoid. Tangent and logistic sigmoidal functions are given below:

$$f_{logistic-sigmoid}(\mathrm{x}) = \frac{1}{1+e^{x}} \tag{2.22}$$

$$f_{tangent-sigmoid}(\mathrm{x}) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{2.23}$$

For representation of calculations, output layer generally has purelin as a linear transfer function or step functions. Step functions can be formulized through the needs of a problem which is usually being clusters or predefined classes. Purelin and step functions are given below:

$$f_{purelin}(\mathrm{x}) = \mathrm{x} \tag{2.24}$$

$$f_{step}(\mathrm{x}) = \begin{cases} 0 & if\ x \in Output1 \\ 1 & if\ x \in Output2 \end{cases} \tag{2.25}$$

In equation (2.25), $Output1$ and $Output2$ are the predefined classes which categorizes the outputs and number of classes can be described by the users according to the problem. All activation functions` graphics are given in Figure 2 for a visual presentation.

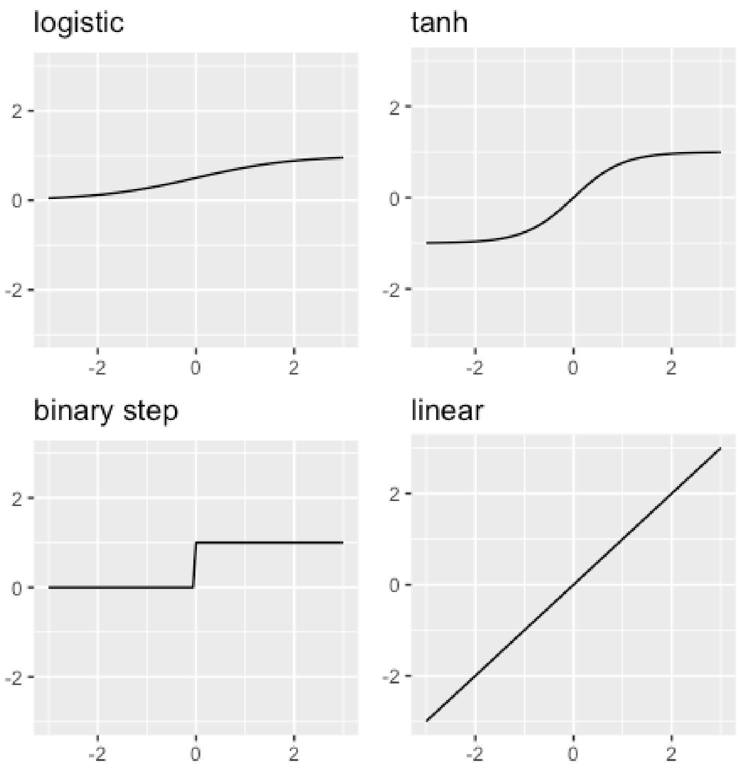

FIGURE 2: Activation functions.

In this study, the tangent-sigmoidal function for hidden layers and the purelin function for the output layer has been set as activation functions of FFNN. Every individual neuron in the networks has its input and output connection. Input and output of a neuron can be calculated as follows:

$$\mathrm{NET}_j = \sum_i w_{ij} f(\mathrm{NET}_i) + w_j \quad (2.26)$$

$$f(\mathrm{NET}_j) \quad (2.27)$$

NET_j is the output of i^{th} neuron, w_{ij} is the weight between i^{th} and j^{th} neuron and w_j is bias. The alteration on the weights will affect the behavior of the connected neuron and therefore the whole network`s behavior.

After calculating the NET_j value with the activation function, $f(\mathrm{NET}_j)$ is the last form of each neuron`s output. This process flows until the very last neuron in the output layer and finally the mathematical form of the trained neural network basically could be presented as follows:

$$Output = f_{purelin}\left(\sum_i w_{ij}{}^{(2)} f_{tangent-sigmoid}\left(w_{ij}{}^{(1)} x_i + w_j{}^{(1)}\right) + w_j{}^{(2)}\right) \quad (2.28)$$

In equation (2.28), inputs x_i given in the network by multiplying with weights between input and hidden layer $w_{ij}{}^{(1)}$ and adding bias term $w_j{}^{(1)}$. After getting into the activation function, the information flows through the hidden layer to the output layer similar to previous layer by multiplying weights $w_{ij}{}^{(2)}$ and adding bias term $w_j{}^{(2)}$.

FFNN handles censored data in a different approach. From the theoretical aspect, due to the incomplete structure of censored data, partially linear regression model estimations have a huge bias, therefore synthetic observations $Z_{i\hat{G}}$ (2.7) should be used as the response variable instead of censored observations y_i (2.1). However, the assumption-free form of FFNN allows it to work with biased or incomplete data. Thus, FFNN doesn't have any limitation as the partially linear regression model about using censored observations y_i (2.1) as the response variable, in other words, target variable of FFNN.

RESULTS

SIMULATION STUDY

To show how methods, perform in estimating the right-censored data, a simulation study is carried out. Performances of FFNN and the semiparametric regression models are compared and discussed. Also, mean squared error (MSE) is used as a performance measure which is calculated as

$$MSE = \sum_{i=1}^{1000} \sum_{j=1}^{n} \left[\frac{(\mathbf{z} - \hat{\mathbf{z}}_{\hat{G}})^2}{n} \right] \quad (3.1)$$

The sample size is determined as *n*=50 and *n*=200, and censoring levels are considered as $C.L. = 10\%$ and 40%.

The dataset (y_i, x_i, t_i) is generated and semi-parametric regression model is obtained as follows:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + f(t_i) + \varepsilon_i, \quad 1 \le i \le n \quad (3.2)$$

where x_{i1} and x_{i2} are obtained from uniform distribution $U[0,1]$, $\boldsymbol{\beta} = (\beta_1, \beta_2)^T = (-2, 0.5)^T$, ε_i's are generated from standard normal distribution $N(0,1)$. Smooth function $f(.)$ is obtained as

$$f(t_i) = 3t_i \sin(t_i), \quad t_i = 6\big((i - 0.5)/n\big) \quad (3.3)$$

To generate the right-censored data, censoring values c_i's are generated from the standard normal distribution with different censoring levels. Values of censoring indicator δ_i's are determined by Bernoulli distribution for 0.10 and 0.40 probabilities which controls the magnitude of the censoring level. The response variable is updated according to censor as follows:

$$z_i = \min(y_i, c_i) \qquad (3.4)$$

Because of the censoring, the response variable is transformed to synthetic data for both SS and FFNN methods.

The input layer`s neuron number has been set as the number of the explanatory variable of each simulated data. Hidden layers` neuron numbers are varying range from 1 to 20 and due to having only one neuron in the output layer, the total number of 400 trained networks are considered for selecting the best architecture in each run. The test set ratio has been decided as 15% for all datasets in the study. Observations in the test set are not a part of the training procedure of networks.

The total number of 100 runs has been evaluated for the $n = 50$ sampled sized and 5 runs for the $n = 200$ sampled size of simulated data. The reason behind the lesser number of runs for larger datasets is the lengthy time needed for accomplishing the task. The results of the simulation study are presented in the following tables and figures.

TABLE 1: Mean squared error values for estimated models of simulated datasets.

Censoring level	10%		40%	
n	**SS**	**FFNN**	**SS**	**FFNN**
50	3.6027	0.0028 (5-3-4-1)	51.0456	0.0755 (5-5-3-1)
200	1.968	0.0094 (5-3-8-1)	27.5817	0.2565 (5-3-8-1)

SS: Smoothing spline; FFNN: Feed forward neural networks.

When Table 1 is inspected, it can be seen that FFNN is the better estimator than SS for the right-censored data. Especially, when the censoring level is high. The cause of the failure of SS has expressed in the next parts. Here, for the lower censoring levels, SS can be interpreted as reasonable but still, the difference between SS and FFNN is large. These results are also supported by Figure 3, Figure 4 also.

In Figure 3 and Figure 4 below, outputs of SS and FFNN are presented for all simulated datasets.

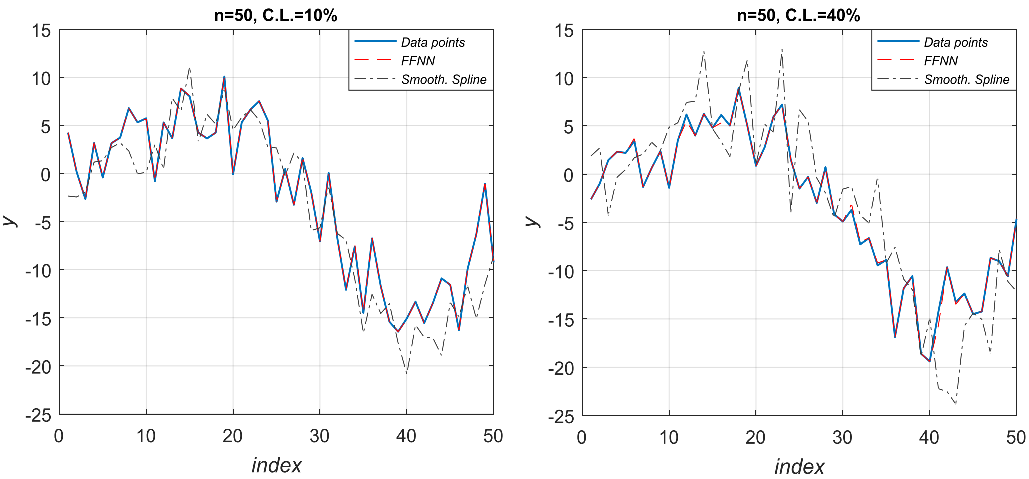

FIGURE 3: Predictions of SS and FFNN for simulated datasets (n=50).

SS: Smoothing spline; FFNN: Feed forward neural networks.

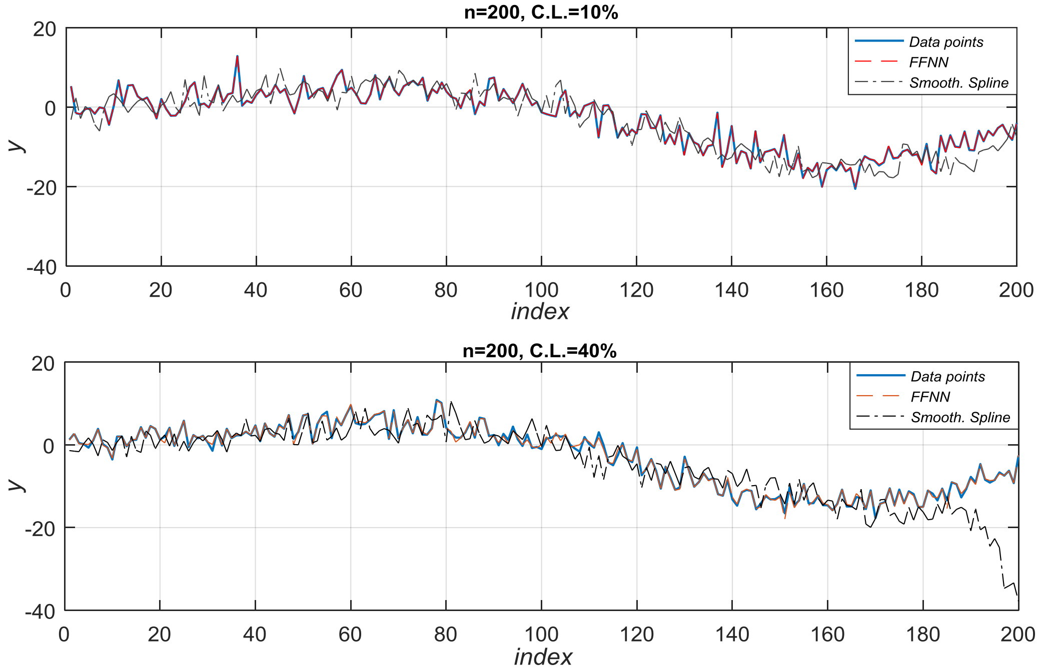

FIGURE 4: Predictions of SS and FFNN for simulated datasets (n=200).

SS: Smoothing spline; FFNN: Feed forward neural networks.

From Figure 3 and Figure 4, FFNN declares its superiority over SS for all censoring levels and sample sizes. Especially in the small sample size, differences between the methods can be seen explicitly. This shows the advantage of FFNN for being able to model small sample sizes. As expected, when the censoring level is getting higher, the performances of the methods decrease. For the larger simulated datasets, FFNN method is much more invulnerable than SS in modeling right-censored data. In Figure 4 that both methods have convincing performances. However, SS uses synthetic data points as knots and therefore it is trying to fit every observation. This situation leads SS to be affected by the censoring mechanism very much which causes it fails on the high censoring levels.

REAL DATA EXAMPLE

The kidney data is used and modeled from a study by McGilchrist and Aisbett.[18] The data set consists of 76 kidney patients and four explanatory variables. The response variable is called as recurrence times of infection (*retime*) and explanatory variables are *age*, *sex* (1=male, 2=female), frailty (*frail*), and disease type (*distype*) coded as GN=0, AN=1 ve PKD=2, 3=other. Note that there is a censoring indicator (1=infection occurs; 0=censored) associated with the response variable. According to the censoring indicator variable, there are 58 censored recurrence time values. For that reason, it is easily seen that the censoring percentage is about 76% and kidney data has a heavy censored.

For this dataset, *frail* is determined as a variable of the nonparametric component, because, it has a nonlinear relationship with the *retime* variable which can be seen clearly in their scatterplot in Figure 5.

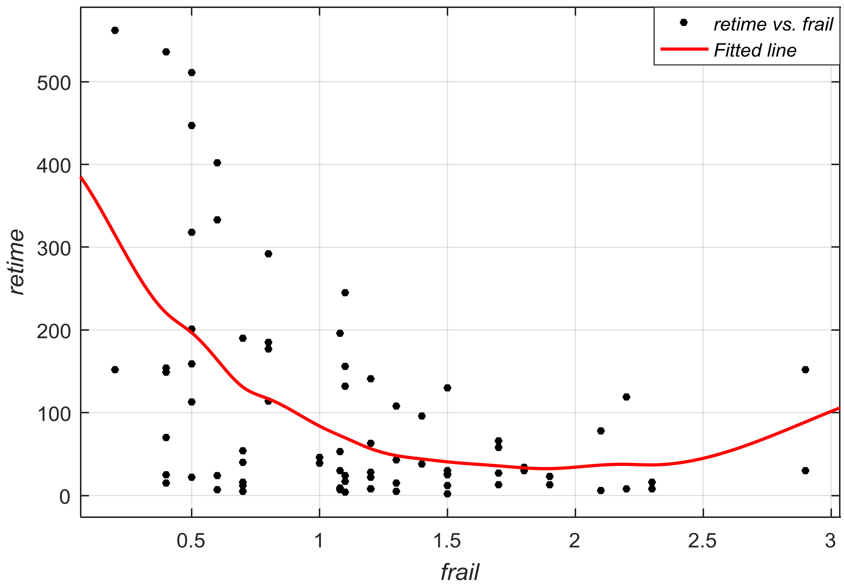

FIGURE 5: Scatter plot and fitted curve for variable *frail* and *retime*.

Figure 5 represents the nonlinear relationship between $frail$ and $retime$. An illustration of this relationship is important for the semiparametric model.

$$retime_i = \beta_1 age_i + \beta_2 sex_i + \beta_3 distype_i + g(frail_i) + \varepsilon_i, i = 1, \dots, 76 \quad (4.1)$$

As indicated in Section 2, $retime_i$ cannot be used as a response variable directly because of the censoring. Consequently, to estimate the model (4.1), the response variable $retime_i$ transformed into synthetic variable $retime_{i\hat{G}}$, as described in equation (2.4).

Unlike the SS, FFNN does not have any assumptions for utilization. This advantage of FFNN allows the researchers to use every single information that can affect the results. In this study, the synthetic variable $retime_{i\hat{G}}$ is also used as an explanatory variable. The architecture of the related FFNN can be seen in Figure 6 below,

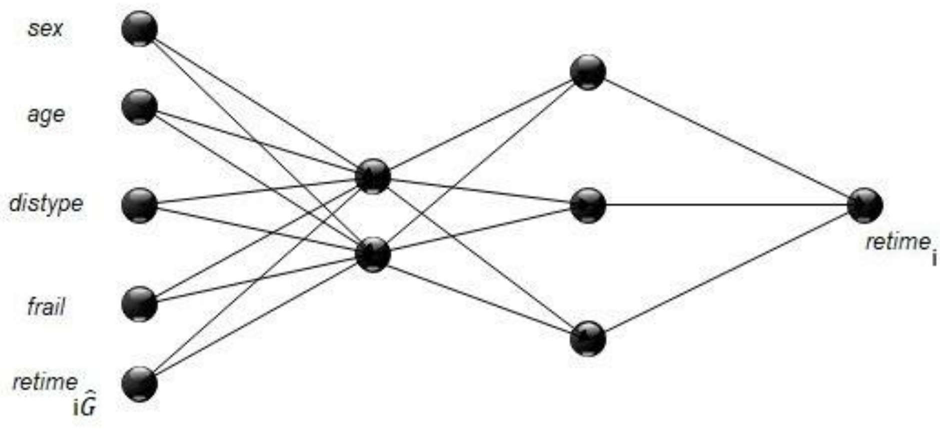

FIGURE 6: Architecture of feed forward neural networks for the real world data.

In this case, the comparative outcomes from both methods SS and FFNN are summarized in Figure 7.

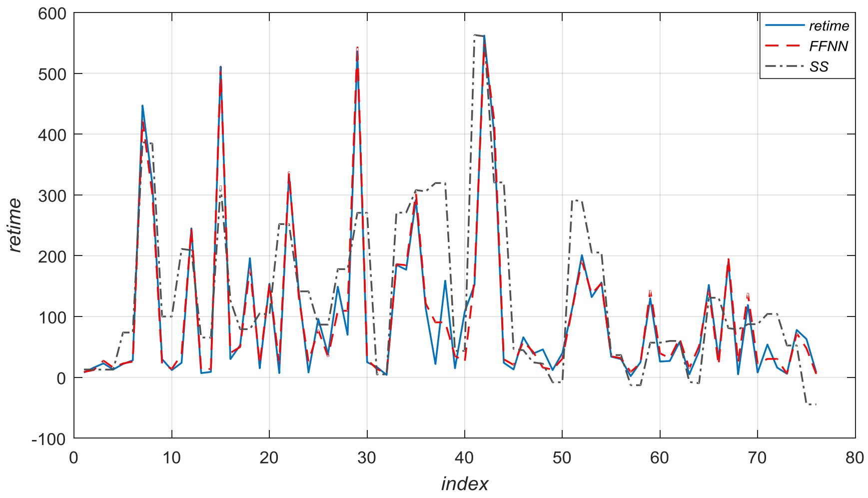

FIGURE 7: Predictions of SS and FFNN for the real world dataset.

SS: Smoothing spline; FFNN: Feed forward neural networks.

As can be seen in Figure 7, FFNN represents the data better than SS. MSE values of FFNN and SS methods are 362.63 and 893.23 respectively. In this study, it must be indicated that FFNN proves itself on its strength when modeling the heavy (76%) censored data. On the other hand, SS shows a reasonable performance on heavily censored data. The reason of the failure of the SS can be expressed as the simulation study with its knot selection.

DISCUSSION

In the literature, these two techniques are compared based on different fields of application by several researchers and each of them reaches different results. For instance, Zain et al. studied the machining performance of the regression and ANN methods and they found that regression reduces the risk slightly more than ANN regarding the results of the empirical study.[19] On the other hand, Adamowski and Karapataki made forecasting for the water-demand by using regression and ANN and according to their results, ANN model based on Levenberg-Marquardt technique gives better prediction performance than regression.[20]

As in Adamowski and Karapataki, ANN shows better performance than the regression model which can be explained by the pointwise estimation success of the ANN. This inference is supported by the simulation and real data example.[20]

CONCLUSION

In this study, two popular estimation methods are introduced to estimate the right-censored data which are FFNN and semi-parametric regression model based on SS respectively. To obtain more accurate results, synthetic data transformation is used. Estimation procedures of the FFNN and SS are discussed and then simulation study and real data work are carried out.

From the results, FFNN shows the best performance for all sample sizes and censoring levels. Especially, under heavily censoring situations, FFNN modeled the data more efficiently than SS. From the results of the real data case, the nonlinear relationship between response variable $retime_i$ and explanatory variable $frail_i$ is illustrated. Also, it can be said that FFNN estimates the right-censored data better than SS same as the simulation study.

Finally, it can be said that the comparison study concludes with the achievement of the FFNN is preferable approach for analyzing censored data. In lower censoring levels, both methods have satisfying results but in heavily censored data, SS cannot perform successfully. Therefore, FFNN will perform better than SS.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and/or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

Idea/Concept: *Ersin Yılmaz, Çağatay Bal;* ***Design:*** *Ersin Yılmaz, Çağatay Bal;* ***Control/Supervision:*** *Ersin Yılmaz, Çağatay Bal;* ***Data Collection and/or Processing:*** *Ersin Yılmaz, Çağatay Bal;* ***Analysis and/or Interpretation:*** *Ersin Yılmaz, Çağatay Bal;* ***Literature Review:*** *Ersin Yılmaz, Çağatay Bal;* ***Writing the Article:*** *Ersin Yılmaz, Çağatay Bal;* ***Critical Review:*** *Ersin Yılmaz.*

REFERENCES

1. Green PJ, Silverman BW. Nonparametric Regression and Generalized Linear Model. 1st ed. London: Chapman & Hall; 1994. [Crossref]
2. Hardle W. Applied Nonparametric Regression. 1st ed. Cambridge: Cambridge University Press; 1990.
3. Koul H, Susarla V, Van Ryzin J. Regression analysis with randomly right-censored data. The Annals of Statistics. 1981;9(6):1276-85. [Crossref]
4. Leurgans S. Linear models, random censoring and synthetic data. Biometrika. 1987;74:301-9. [Crossref]
5. Qin G, Jing B. Asymptotic properties for estimation of partial linear models with censored data. Journal of Statistical Planning and Inference. 2000;84(1-2):95-110. [Crossref]
6. Orbe J, Ferreira E, Nú-ez-Antón V. Censored partial regression. Biostatistics. 2003;4(1):109-21. [Crossref] [PubMed]
7. Aydın D, Yılmaz E. Modified estimators in semi-parametric regression models with right-censored data. Journal of Statistical Computation and Simulation. 2018;88:1470-98. [Crossref]
8. Faraggi D, Simon RA. Neural network model for survival data. Statistics in Medicine. 1995;14:73-82. [Crossref] [PubMed]
9. Biganzoli E, Boracchi P, Mariani L, Marubini E. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. Stat Med. 1998;17(10):1169-86. [Crossref] [PubMed]
10. Biganzoli E, Boracchi P, Marubini E. A general framework for neural network models on censored survival data. Neural Netw. 2002;15(2):209-18. [Crossref] [PubMed]
11. Azadeh A, Keramati A, Tolouei H, Parvari R, Pashapour S. Estimation and optimization of right-censored data in survival analysis by neural network. International Journal of Business Information Systems. 2013;14:322-34. [Crossref]
12. Kalderstam J, Edén P, Bendahl PO, Strand C, Fernö M, Ohlsson M. Training artificial neural networks directly on the concordance index for censored data using genetic algorithms. Artif Intell Med. 2013;58(2):125-32. [Crossref] [PubMed]
13. Stute W. Nonlinear censored regression. Statistica Sinica. 1999;9:1089-102. [Link]
14. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association. 1958;53(282):457-81. [Crossref]
15. Werbos PJ. Beyond regression: new tools for prediction and analysis in the behavioral sciences [PhD thesis]. Cambridge: Harvard University; 1974. Erişim Tarih: 2022 [Link]
16. Nagl S. Neural network models of protein domain evolution. Hyle. 2000;6:143-59. [Link]
17. Kadılar C, Şimşek M, Aladağ CH. Forecasting the exchange rate series with ANN: the case of Turkey. Istanbul University Journal of Econometrics and Statistics. 2009;9:17-29. [Link]
18. McGilchrist CA, Aisbett CW. Regression with frailty in survival analysis. Biometrics. 1991;47(2):461-6. [Crossref] [PubMed]
19. Zain AM, Haron H, Qasem SN, Sharif S. Regression and ANN models for estimating minimum value of machining performance. Applied Mathematical Modelling. 2012;36(4):1477-92. [Crossref]
20. Adamowski J, Karapataki C. Comparison of multivariate regression and artificial neural networks for peak urban water-demand forecasting: evaluation of different ANN learning algorithms. Journal of Hydrologic Engineering. 2010;15(10):729-43. [Crossref]