İSTATİSTİK: JOURNAL OF THE TURKISH STATISTICAL ASSOCIATION
Vol. 11, No. 1-2, January-July 2018, pp. 1–11
ISSN 1300-4077 | 18 | 1-2 | 1 | 11

İSTATİSTİK

A MODELLING APPROACH TO INCREASE THE EXPLAINED RISK IN THE PROPORTIONAL HAZARDS REGRESSION

Deniz Inan*
Department of Statistics,
Marmara University,
34730, Istanbul, Turkey

Oykum Esra Askin
Department of Statistics,
Yildiz Technical University,
34220, Istanbul, Turkey

Abstract: In this study, a modelling strategy is developed to obtain more information from censored observations. By the proposed approach, uncensored observations are clustered using a fuzzy c-means algorithm and the degrees to which censored observations are members of these clusters are determined. Censored observations are weighted based on their membership values and the distances between the censoring time and the time components of the cluster centres. Further, simulation studies are performed to characterize the performance of the proposed approach based on the explained risk measure.

Key words: Censored data, Explained risk, Fuzzy c-means algorithm, Proportional hazards regression models, Survival analysis
History: Submitted: 29 June 2017; Revised: 21 December 2017; Accepted: 20 January 2018

1. Introduction

Survival analysis is an important statistical technique that allows researchers to investigate the risk factors influencing individual survival time. It is different from other statistical techniques with respect to the data censoring. Unlike classical techniques, such as logistic regression, survival analysis does not exclude incomplete observations; instead, it uses both censored and uncensored observations in order to increase the amount of information obtained for the subsequent modelling process. Censored observations are found when some information is known about the survival times of individuals but not the exact survival times. There are different censoring types used in survival analysis. But, in this study, right censoring is considered. In general, right censoring may occur under three conditions: if a unit does not fail before the study ends, if a unit is lost to follow-up during the study, or if a unit withdraws from the study because of a failure (in cases when failure is not the event of interest) [10]. Suppose that there are n observations in the study and t_i denotes the observed failure time of i^{th} observation. Under a right censoring scheme, $t_i = \min(c_i, t_i^*)$ where t_i^* is the failure time and c_i is the censoring time. Here, t_i^* and c_i are independent random variables. The observed censoring indicator δ_i is equal to 1 if $t_i^* < c_i$, and 0 otherwise.

One way to investigate the relative risk is to model the data using the proportional hazards (PH) approach that was introduced in [4]. According to information observed from the distributional form of the failure times, either a fully parametric or a semi-parametric PH model can be used. The PH model has the form:

* Corresponding author. E-mail address:denizlukuslu@marmara.edu.tr

$$h(t) = h_0(t) \exp(\beta' X), \tag{1.1}$$

where $h_0(.)$ is the baseline hazard function, X is the vector of covariates (risk factors) $X' = (X_1, ..., X_p)$ and β is a vector of regression parameters which shows the relation between risk factors and risk of the failure $\beta' = (\beta_1, ..., \beta_p)$.

Determining the failure risks for different units that have different covariate profiles is important for making statistical inferences. The level of risk explained by the PH model is directly proportional to the amount of information obtained from the data. Therefore, it is important to incorporate censored observations into the model even if they do not contain complete information. While some of the censored observations contain almost as much information as an uncensored observation, some of them contain very little information. Therefore, their contributions to the model may vary.

The aim of this study is to weigh the censored observations according to the amount of information they contain and, in this way, further increase the amount of information obtained from the censored observations.

In homogeneous populations, it is expected that observations with similar covariate profiles have failure times that are close together. Assume that there are three observations with the same covariate values and that one of them is uncensored while the others are right censored with different censoring times. It is clear that a censored observation that stays in the study longer (in other words, whose censoring time is close to the failure time of an uncensored observation), provides more information to the model. Therefore, in this case, it can be said that the distance between the censoring time of a censored observations and the failure time of an uncensored one can be considered a measure of the contribution of the censored observation to the model.

Of course, in real life problems, there are covariate profiles are often similar but not exactly the same. In this context, in order to evaluate the contribution of a censored observation, it is important to identify the cluster of uncensored observations which are the most similar to that observation. It is also important to quantify the degree of similarity them.

For this purpose, a fuzzy c-means clustering algorithm (FCM) is used as an auxiliary tool. What makes FCM different is that it does not decide the absolute membership of a data point to a given cluster; instead, it calculates the degree of membership that a data point will belong to that cluster. First, clusters of the most similar uncensored observations are defined. Then, for each censored observation, the most similar cluster is identified and the membership value (which is used as a measure of the degree of similarity) with respect to this cluster is determined based on the membership formula of the FCM.

The remainder of this article is organized as follows. Section 2 describes the general concept of the FCM. Section 3 outlines the proposed modelling strategy. Section 4 describes the performance of proposed approach based on simulation studies. Section 5 describes an application to real data to verify the efficiency of the proposed approach. Finally, Section 6 presents some concluding remarks.

2. Fuzzy c-means algorithm

FCM was proposed by Dunn [5] in 1973 and was later improved upon by Bezdek [1] in 1981; since then, it has proven to be one of the most important fuzzy clustering algorithms. Given a set of data, clustering techniques are used to partition the data points into several groups such that the degree of association within each group is high and that between the data points in different groups is low. Classical crisp clustering techniques result in crisp partitions such that each data point may belong to only one cluster. In contrast, fuzzy clustering results in fuzzy partitions such that any data point may belong to more than one group. Each cluster is associated with a membership function that expresses the degree to which the individual data points belong to the cluster. Of the various fuzzy clustering methods that have been proposed, FCM remains dominant in the literature as it has been successfully applied in both academia and industry.

FCM performs clustering by iteratively searching for a set of fuzzy clusters and the associated cluster centres that represent the structure of the data as closely as possible. The algorithm requires the user to specify the number of clusters to be formed from the dataset, c, and partitions the data into c fuzzy clusters by minimising the sum of squared error objective function (given in Equation 2.1) within each group.

$$J_m(U,V) = \sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}^{m} \left\| y_i - v_j \right\|^{2}, \tag{2.1}$$

subject to

$$0 \le u_{ij} \le 1, \quad \forall i,j$$
$$0 \le \sum_{i=1}^{n} u_{ij} \le 1, \quad \forall j$$
$$\sum_{j=1}^{c} u_{ij} = 1, \quad \forall i$$

In the objective function, $m \in [1,\infty)$ represents the level of fuzziness of the cluster and is referred to as the fuzziness parameter; values m in the range of 1.5-3 are effective in many applications [2]. $Y = \{y_1, y_2, ...y_n\}$ represents the observations, n represents the number of observations and c represents the number of clusters. $u_{ij} \in [0,1]$ is the membership value of y_i in j^{th} cluster and v_j represents the cluster center of j^{th} cluster. Here, $\|.\|$ is the Euclidean distance used in objective function.

As understood, this method allows an observation to belong to more than one cluster. The values of u_{ij} and v_j are updated in each iteration of algorithm. The FCM algorithm includes the following steps:

Step 1 Initialize membership value matrix $U = [u_{ij}]$ with the initial value $U^{(0)}$, randomly.

Step 2 Set the values of c, m and the termination criteria ε.

Step 3 Compute v_j which is given below:

$$v_j = \frac{\sum_{i=1}^{n} u_{ij}^{m} y_i}{\sum_{i=1}^{n} u_{ij}^{m}}. \tag{2.2}$$

Step 4 Compute and update U with following Equation:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\left\| y_i - v_j \right\|}{\left\| y_i - v_k \right\|} \right)^{\frac{2}{m-1}}}. \tag{2.3}$$

Step 5 If $\left\| U^{(b)} - U^{(b-1)} \right\| \le \varepsilon$ $(b = 0,1,...)$ stop, otherwise return to Step 3.

3. Proposed method

First, uncensored observations are clustered using FCM based on the corresponding vector of covariates and failure times. Each uncensored observation is assigned to a class that is linked with the highest membership value. Only uncensored observations, which contain full sets of information, are used here in order to accurately define the time components of the cluster centers.

Then, the membership values of each of the censored observations to each of the clusters are determined using only the corresponding vectors of covariates and applying the FCM membership value formula given in Equation 2.3. For each censored observation, the cluster associated with the greatest membership value is considered to be the cluster for that observation. In this way, most of

the uncensored observations with similar covariate structures can be combined to form a censored observation.

Finally, an additional weight covariate (W_{cov}) is calculated in order to increase the information obtained from the censored observations based on the following two quantities:

1. The maximum membership values of the censored observations
2. The distances between the censoring times, t^{cen}, and the time components of the corresponding cluster centres of the censored observations.

The detailed steps of the proposed method are given below:

Step 1 The number of clusters, the fuzziness parameter and the termination criteria are set.

Step 2 The cluster centres are defined based on the uncensored observations by applying the FCM.

$$v_j = (t_j^*, x_{(1,j)}^*, x_{(2,j)}^*, ..., x_{(p,j)}^*) \quad j = 1, 2, ..., c$$

Cluster centres do not have to be one of the observations so they are denoted as (*).

Step 3 For each of the censored observations, membership values to all clusters are computed using the following equation:

$$u_{hj} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\left\| y_h - v_j^{\#} \right\|}{\left\| y_h - v_k^{\#} \right\|} \right)^{\frac{2}{m-1}}},$$

where $v_j^{\#} = (x_{(1,j)}^*, x_{(2,j)}^*, ..., x_{(p,j)}^*)$, $y_h = (x_{(1,h)}, x_{(2,h)}, ..., x_{(p,h)})$, $j = 1, 2, ..., c$ and $h = 1, 2, ...,$"number of censored observations".

Step 4 Maximum membership values (u_h^*) and the corresponding cluster centers $v^* = (t^*, x_1^*, x_2^*, ..., x_p^*)$ are determined with the condition of $t_h^{cen} < t^*$.

Step 5 Distance between censoring times and the time components of corresponding cluster centers are determined as:

$$d_h^* = t^* - t_h^{cen}$$

Step 6 Weight covariate is defined as below:

$$W_{\text{cov}} = \begin{cases} 0, & \text{for uncensored observations} \\ \frac{u_h^*}{d_h^* / \max(t)} & h = 1, 2, \ldots, \text{" number of censored observations"}. \end{cases}$$

where $\max(t)$ is the maximum value of t.

Step 7 W_{cov} is used in modelling process as an additional covariate.

4. Simulation studies

To illustrate the performance of the proposed method in terms of the explained risk measures, two scenarios are considered. For both scenarios, the baseline hazard function is assumed to have a Weibull distribution with the shape parameter $p = 0.25$ and the scale parameter λ is re-parametrized as $1/\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3)$. The following form of the PH model is used:

$$h(t) = pt^{p-1}(\exp(-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3)))^p$$

In the first scenario, three covariates are randomly generated from the standard normal distribution and included in the model. The regression parameters are set to be $\beta_1 = \beta_2 = \beta_3 = -1$. The censoring distribution is considered to be a Weibull distribution to attain the desired censoring rate of nearly 45%, which represents heavy censoring. In the second scenario, three covariates are included as in the first scenario but, in contrast, two of the covariates are generated from Bernoulli distributions with success probabilities of 0.5 and 0.7 and third is generated from a standard normal distribution. The regression parameters are assumed to be $\beta_1 = -2, \beta_2 = -0.2, \beta_3 = -1$. The censoring distribution is considered to be a Weibull distribution and the censoring rate is set as 25%, which represents medium censoring.

Training and testing data sets were generated randomly for both of scenarios with various sample sizes: $(N_{train}, N_{test}) = (1000, 500)$, $(N_{train}, N_{test}) = (500, 250)$ and $(N_{train}, N_{test}) = (300, 100)$. Each

of these sample size combinations were tested with different numbers of clusters: $c=(5,8,10,15)$, $(4,5,8,10)$ and $(2,3,4,5)$, respectively. Thus, twelve different cases were examined for each scenario.

Simulations were repeated 100 times for each case. The standard model and the proposed PH model were fitted to training data sets. Then, to test the significance of the W_{cov}, the likelihood ratio (LR) test statistic (which has a Chi-squared distribution with p degrees of freedom where p denotes the number of predictors being assessed) was used. Thus, with a significance level of $\alpha=0.05$, the critical value was 3.8.

Within the context of the PH model, the explained risk is used to quantify the ability to determine a patient's risk based on his or her covariate profile. The coefficient of determination, R^2, is a standard measure of explained risk in a normal linear model with uncensored data.However, due to the difficulty of attaining an R^2 type measure with censored data and a PH model, different measures of explained risk have been constructed for survival data [3], such as R^2_{LR}. In this study, the standard PH model and the proposed model are applied to the testing data sets and the prediction performances of the two approaches are compared in terms of the R^2_{LR} values, which is defined as follows [9]:

$$R^2_{LR}=1-\left(\frac{l(0)}{l(\widehat{\beta})}\right)^{2/d}, \tag{4.1}$$

where $l(\widehat{\beta})$ and $l(0)$ are the values of likelihood function with and without covariates, respectively and d is the number of uncensored observations (failures). As shown in equation, R^2_{LR} is based on the LR statistic [7].

In order to show the results of 24 different cases less complicated, the mean values of LR and R^2_{LR} is reported. Results are given in Table 1 and Table 2 for scenario 1 and scenario 2, respectively.

As seen in Table 1 and Table 2, mean value of LR test statistics is very high according to the critical χ^2 value of 3.84 in all cases. Besides, in all cases, R^2_{LR} values of proposed method is better than the standard PH which means that the model constructed using the proposed method gives better predictions on the test data.

TABLE 1. Simulation results for scenario 1

	Scenario	Number of cluster	mean(LR)	$\bar{R}^2_{LR}$ (standard PH)	$\bar{R}^2_{LR}$ (proposed model)
Scenario 1: 40% censored	$(N_{train},N_{test})=(300,100)$	2	33.43	0.72	0.75
		3	90.71	0.71	0.78
		4	117.94	0.70	0.79
		5	144.22	0.72	0.81
	$(N_{train},N_{test})=(500,250)$	4	195.85	0.72	0.79
		5	263.42	0.73	0.82
		8	322.57	0.72	0.83
		10	372.77	0.73	0.85
	$(N_{train},N_{test})=(1000,500)$	5	539.07	0.73	0.82
		8	708.00	0.73	0.85
		10	752.12	0.73	0.85
		15	820.55	0.73	0.86

TABLE 2. Simulation results for scenario 2

	Scenario	Number of cluster	mean(LR)	$\bar{R}^2_{LR}$ (standard PH)	$\bar{R}^2_{LR}$ (proposed model)
Scenario 2: 25% censored	$(N_{train}, N_{test}) = (300, 100)$	2	26.48	0.56	0.58
		3	79.82	0.57	0.62
		4	114.92	0.58	0.65
		5	128.64	0.58	0.65
	$(N_{train}, N_{test}) = (500, 250)$	4	157.06	0.58	0.63
		5	230.82	0.57	0.65
		8	298.38	0.58	0.68
		10	308.34	0.57	0.67
	$(N_{train}, N_{test}) = (1000, 500)$	5	513.47	0.67	0.80
		8	600.05	0.66	0.81
		10	650.88	0.66	0.83
		15	711.73	0.71	0.88

To further illustrate the superiority of proposed method compared to the standard PH method, the case $(N_{train}, N_{test}, c) = (1000, 500, 15)$ is randomly chosen from scenario 1. Figure 1 shows the LR values for the training sets for each repeat, the R^2_{LR} values for testing data sets for each repeat and survival curves of a randomly chosen iteration obtained from applying the standard, proposed and Kaplan-Meier (KM) methods to the testing data set.

Figure 1a shows that all LR values obtained with the proposed method are above the critical value of 3.84, which means that the W_{cov} is significant for each repeat. Further, Figure 1b shows that the proposed method produces better R^2_{LR} values when applied to testing data sets, which indicates that adding W_{cov} improves the model fitting. Figure 1c illustrates survival curves obtained from; standard, proposed and Kaplan Meier methods for testing data set of a randomly chosen iteration. It is clearly seen in Figure 1c, survival curves of proposed method are closer to KM fits when compared with the standard PH method's. This confirms that the proposed method performs better in terms of the explained risk.

To confirm these findings, this investigation was repeated for Scenario 2 with the case $(N_{train}, N_{test}, c) = (500, 250, 10)$. The results shown in Figure 2a, 2b and 2c depict the LR test values from the training set in each repeat, the R^2_{LR} values from the testing data sets in each repeat and the survival curves from the testing data set for a randomly chosen iteration, respectively. The results for this case are consistent with those of the other cases tested.

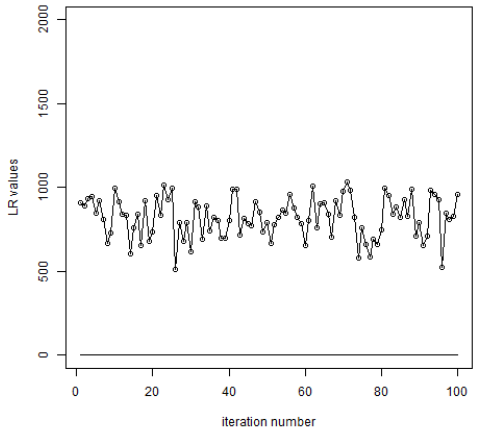

Figure 1a:Change of LR test values

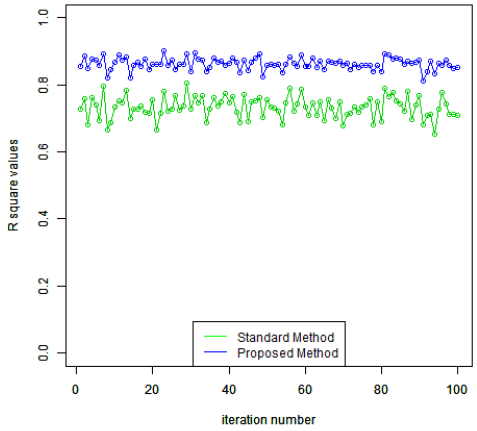

Figure 1b: Change of R^2 values

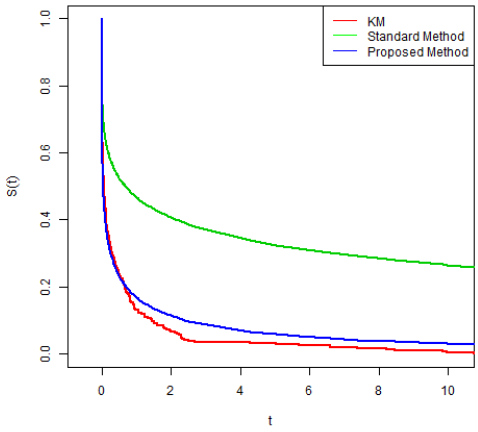

Figure 1c: Survival curves
Figure 1. The case $(N_{train}, N_{test}, c) = (1000, 500, 15)$

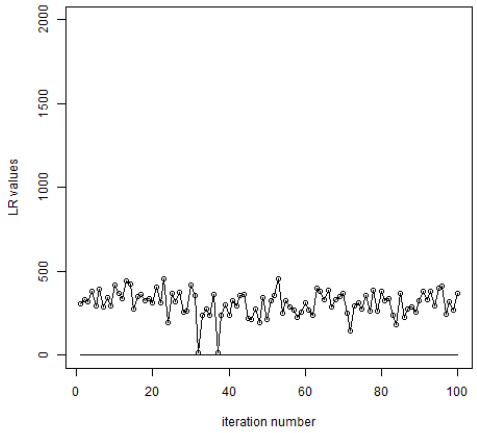

Figure 2a:Change of LR test values

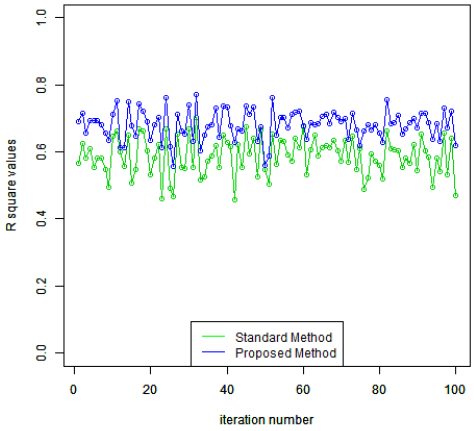

Figure 2b: Change of R^2 values

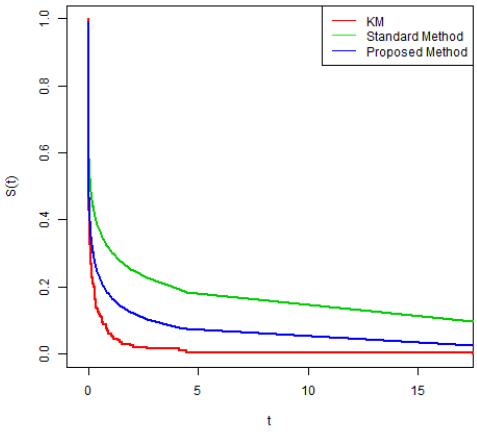

Figure 2c: Survival curves
Figure 2. The case $(N_{train}, N_{test}, c) = (500, 250, 10)$

5. Real data application

The data from the Worcester Heart Attack Study [6, 8] has been widely used in survival studies. Data concerns the survival times of 500 patients having their first attack. Also 22 different covariates belong to patients were reported such as age, gender, body mass index (BMI), length of hospital stay and initial systolic blood pressure in order to build survival models. Censored rate of data set is nearly 56.5%

In this study, three primary risk factors such as age, gender and BMI are included in the model. The hazard function at time $t > 0$ takes the form:

$$h(t) = pt^{p-1}(\exp(-(\beta_0 + \beta_1 age + \beta_2 gender + \beta_3 BMI)))^p \tag{5.1}$$

The data set was randomly divided as training and testing data sets. 400 of the observations are used as training and uncensored observations of the remaining 100 are used as testing. Standard and proposed PH models are fitted on training data set. To perform FCM procedure, firstly number

of clusters is chosen as 3. The estimation results obtained from the standard and proposed PH models are shown in Table 3. As seen in table, the proposed method provides a higher logarithmic likelihood value. Besides, proposed method gives slightly lower standard errors for coefficients when compared with standard method. Further, from the testing data set, the R^2_{LR} attained by the standard method was 0.70 while that attained with proposed method was 0.84.

TABLE 3. Real data results for the number of clusters 3

	PH Method		Proposed Method	
	Coefficient	S.E.	Coefficient	S.E.
β_0	12.962	1.532	11.676	1.453
β_1	0.106	0.032	0.066	0.031
β_2	-0.103	0.014	-0.082	0.013
β_3	0.168	0.296	0.013	0.292
p	1.893	0.067	1.885	0.065
Log.Lik.	-1396.3		-1326.9	

The case where c=5 was also investigated and the results are given in Table 4. The results show that the proposed method provides a higher logarithmic likelihood value when c=5 than when c=3. Also, from the testing data set, R^2_{LR} attained with the proposed method with c=5 was 0.91, which is greater than that attained with c=3. Thus, it can be concluded that when larger cluster numbers are used, the logarithmic likelihood value is unchanged. Therefore, five clusters is considered suitable for the Worcester Heart Attack Study data.

TABLE 4. Real data results for the number of clusters 5

	PH Method		Proposed Method	
	Coefficient	S.E.	Coefficient	S.E.
β_0	12.962	1.535	5.400	1.224
β_1	0.105	0.032	0.057	0.026
β_2	-0.102	0.014	-0.014	0.010
β_3	0.168	0.296	0.291	0.235
p	1.898	0.067	1.548	0.062
Log.Lik.	-1396.3		-1213.2	

Figure 3 and Figure 4 show the survival curves based on the testing data set that were obtained using the proposed, standard PH and KM methods in the first and second cases, respectively. The survival curves obtained by the proposed method are closer to the KM survival curve in both cases, further illustrating the performance of the proposed method. As expected, in the second case, the survival curve obtained by the proposed method was much closer to KM survival curve than that obtained by the standard method.

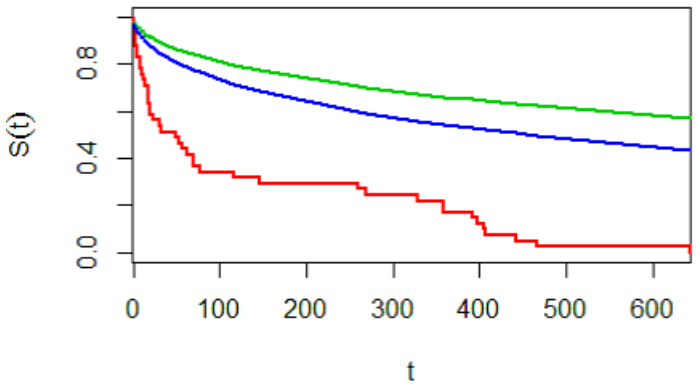

Figure 3: Survival Curves for Number of Clusters is 3

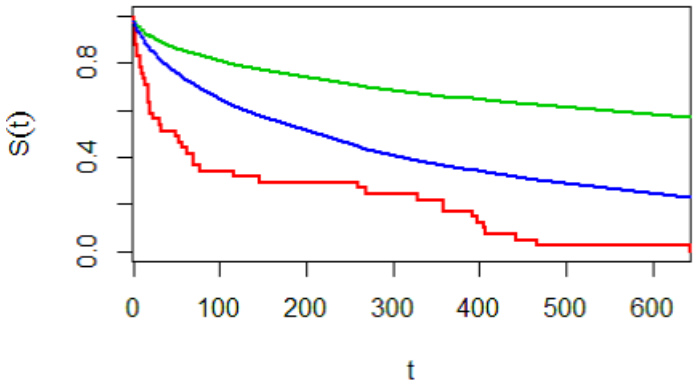

Figure 4: Survival Curves for Number of Clusters is 5

6. Conclusion

In this study, a novel approach to increase the prediction performance of PH models was proposed. This method involves increasing the information obtained from the right-censored data. The performance of the proposed method was first investigated through simulation studies. The results of the simulations showed that increasing the number of clusters results in increases in both the average likelihood value obtained with the training data set and the mean R^2_{LR} value obtained with the test data sets. This result was expected because increasing the number of clusters also increases the amount of information included in the model. Although only randomly selected cases were presented here, it was observed that the proposed method performs better than the classical method at each iteration under all cases.

In addition, the proposed method was applied to real data with different numbers of clusters. In both cases, it was observed that the standard error in the parameter estimates attained by the proposed method was slightly lower than that with the classical method, further confirming the positive effect of increasing the number of clusters. However, as the number of clusters was further

increased, the performance of the method reached a plateau. In this way, the optimal number of clusters for the data set could be determined.

Considering the results of both the simulation studies and the application to real data, the use of the proposed method increases the prediction performance of the PH model. The proposed approach gives good results, particularly when the data is highly censored.

References

[1] Bezdek, J.C. (1971). *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York.

[2] Bezdek, J.C., Ehrlich R. and Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers and Geosciences*, 10(1984), 191-203.

[3] Choodari-Oskooei, B., Royston P. and Parmar, M. (2012). A simulation study of predictive ability measures in a survival model 1: explained variation measures. *Statistics in Medicine*, 31, 2627-2643.

[4] Cox, D. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society B*, 34, 187-220.

[5] Dunn, J.C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated cluster. *Journal of Cybernetics*, 3, 32-57.

[6] Goldberg, R.J., Gore, J.M., Alpert J.S. and Dalen, J.E. (1986). Recent changes in attack and survival rates of acute myocardial infarction (1975 through 1981): the Worcester heart attack study. *Journal of the American Medical Association*, 255, 2774-2779.

[7] Heller, G. (2012). A measure of explained risk in the proportional hazards model. *Biostatistics*, 13, 315-325.

[8] Hosmer Jr., D.W., Lemeshow, S. and May, S., 2008. *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*, Wiley, Hoboken.

[9] Kent, J.T. (1983). Information gain and a general measure of correlation. *Biometrika*, 70, 163-173.

[10] Kleinbaum, D. and Klein, M. (2005). *Survival Analysis: A Self-Learning Text*, Springer, New York.