Survival Model for Identifying the Influencing Factors of Heart Disease

R. Mahalakshmi[1*], K. Senthamarai Kannan[2]
***Corresponding Author:-** R. Mahalakshmi
[1,2] Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli – 627 012.
[1]Email : mahalakshmirajendran@gmail.com,
[1] Email: senkannan2002@gmail.com

Abstract
The early prognosis of cardiovascular disease and which factor decreases the life span of the heart failure patients at most will help us to decreases the deaths due to heart failure. The main objective of this paper is to fit a cox regression model for identify the factors which influence the life span of heart disease people. Also, we used survival analysis techniques to find out which factor decreases the life span of the heart failure patients for this dataset.

Keywords: Survival Model, CVD and Cox Regression
DOI Number: 10.14704/nq.2022.20.9.NQ44695 **Neuro Quantology 2022; 20(9):5961-5964**

Introduction:

Cardiovascular disease cause death globally, more people die annually from cardiovascular disease than from any other cause. An estimated 17.9 million people died from CVDs in 2016, representing 31% of all global deaths. Of these deaths, 85% are due to heart attack and stroke. Over three quarters of CVD deaths take place in low- and middle-income countries. Out of the 17 million premature deaths (under the age of 70) due to non-communicable diseases in 2015, 82% are in low- and middle-income.

In total, the following thirteen variables are considered for the analysis. Three predictors are basic information of the patients i.e., age, gender and smoker. One factor is time which indicate the follow up time of a heart disease patient and one variable is the death event which indicates whether the patient died during the follow up time or not. Other eight predictors are biological information of the patients. Recorded information about two hundred and ninety nine patients are considered.

At first, numerical information pertaining to all the study variables are described using descriptive statistical measures. Association of the target variable with each predictor is analyzed applying Chi-square test and degree of association is calculated using Cramer's V co-efficient. The predictors having high influence upon the target variable are considered for construction of multiple logistic regression model. Significance of the constructed model is tested applying likelihood ratio test. Significance of each predictor is tested based on Wald's statistic. Finally, a statistically significant multiple logistic regression model is obtained with six significant predictors. Efficiency and classification potency of the model are verified using Hosmer-Lemeshow test, area under the curve, and classification potency. For each variable Kaplan –Meier analysis was carried out and then cox regression model was fitted for the given information. Using this result the factors which influence the lifetime of a heart disease patient was found.

CVD risk models are based on single risk factor which cannot realize the influence of multiple factors simultaneously. Risk models using statistical regression methods prefer to use classic risk factors such as age, smoking, diabetes, sex, high blood pressure, and total cholesterol to estimate the risk score. Studies applying data mining or machine learning techniques for the CVD risk estimations cannot provide an absolute risk estimation, although some of these models tried to incorporate novel predictors in the risk models. This research aims to identify the novel risk factors for CVD detection by conventional predictors and then

enhance the risk estimation by developing a multiple-variable-based risk prediction model that targets the 5-year and 10-year CVD events.

Methodology:

Assume we have n independent observations, each containing information on the length of time a subject observed, whether the observation was a survival time or a right censored and a single covariate whose value is determined at the time observation begins and remains at that value throughout the follow-up of the subject. Covariate whose values are fixed at the value measured at the beginning of follow-up are most commonly ensured in practice. Denote the triplet of observed time, covariate and censoring variables as (t_i, x_i, c_i), i=1,2,…,n.

The first step in maximum likelihood estimation is to create the specific likelihood function to be maximized. In simplest terms, the likelihood function is an expression that yields a quantity similar to the probability of the observed data under the model. Suppose that the distribution of survival time for a subject with a single covariate can be described by the cumulative distribution function F (t, β, x). The survival function is obtained from the cumulative distribution function and is defined as

$$S(t, \beta, x) = 1 - F(t, \beta, x) \quad \ldots (1)$$

To create the likelihood function we need a function that we think of, for the moment, as giving the "probability" that the survival time is exactly t. this function is derived mathematically from the distribution function and is called the density function. We denote the density function corresponding to F (t, β, x) as f (t, β, x).

We construct the actual likelihood function by considering the contribution of the triplets (t, 1, x) and (t, 0, x) separately. In the case of the triplet (t, 1, x) we know that the survival time was exactly t. Thus the contribution to the likelihood for this triplet is the "probability" that a subject with covariate value x dies from that disease of interest at time t units. This is given by the value of the density function f (t, β, x). For the triplet (t, 0, x) we know that the survival time was at least t. Thus, the contribution to the likelihood function of this triplet is the probability that a subject with covariate value x survives at least t time units. This probability is given by the survival function S (t, β, x) for censored observations. Under the assumption of independent observations , the full likelihood function is obtained by multiplying the respective contribution of the observed triplets, a value of f (t, β, x) for a non-censored observation and a value of S (t, β, x) for censored observations. In general, a concise way to denote the contribution of each triplet to likelihood is the expression $x\ [S(t, \beta, x)]^{1-c}$ where c=0 or 1. … (2)

Because the observations are assumed to be independent, the likelihood function is the product of the expression over the entire sample is $l(\beta) = \prod_{i=1}^{n} \{ [f(t, \beta, x)]^{c_i} \times [S(t, \beta, x)]^{1-c_i} \}$ …(3)

To obtain the maximum likelihood with respect to the parameter of interest, β , we maximize the log-likelihood function

$$L(\beta) = \sum_{i=1}^{n} \{C_i \ln [f(t, \beta, x)] + (1 - C_i) \ln [S(t, \beta, x)]\} \quad \ldots (4)$$

Because the log function is monotone, the maximum of (2) and (3) occur at the same value of β; however, maximizing (3) is computationally simpler than maximizing (2). The procedure to obtain the values of the MLE involves taking the derivative of L (β) with respect to β, the unknown parameter, setting these equations equal to zero, and solving for β. An application of methods from calculus shows that the density function is the product of the hazard function and the survival function, yielding the expression f (t, β, x) = h (t, β, x) x S (t, β, x). … (5)

Substituting (5) into the log likelihood equation in (4) and simplifying using the form of the proportional hazards model in (1) yields

$$L(\beta) = \sum_{i=1}^{n} \{C_i \ln[h_0(t)] + C_i x_i \beta + (e^{x_i \beta} \ln [S_0(t_0)]\} \quad \ldots (6)$$

Full maximum likelihood requires that to maximize (6) with respect to the unknown parameter of the interest, β, and the unspecified baseline hazard and survival functions. As discussed in Kalbfleisch and prentice (2002) it is not possible to use the log-likelihood function in (6). Therefore, the proportional hazards model in (6) is chosen to avoid having to specify explicitly the error component of the model.

Cox (1972) proposed using an expression called "partial likelihood function" that depends only on the parameter of interest. He speculated that the resulting parameter estimators from the

partial likelihood function would have the same distributional properties as full maximum likelihood estimators. The partial likelihood is given by the expression $l_p(\beta) =$

$$\prod_{i=1}^{n}\left[\frac{e^{x_i\beta}}{\sum_{j\in R(t_i)} e^{x_i\beta}}\right]^{c_i} \quad \ldots (7)$$

The expression in (7) assumes that there are no tied times, and it is often modified to exclude term when $c_i = 0$, yielding

$$l_p(\beta) = \prod_{i=1}^{m}\left[\frac{e^{x_i\beta}}{\sum_{j\in R(t_i)} e^{x_i\beta}}\right] \quad \ldots (8)$$

Where the product is over the m distinct ordered survival times and $x_{(i)}$ denotes the value of the covariate for the subjects with ordered survival time $t_{(i)}$. The log partial likelihood function is

$$L_p(\beta) = \sum_{i=1}^{m}\left\{x_{(i)}\beta - \ln\left[\sum_{j\epsilon\epsilon R(t_{(i)})} e^{x_j\beta}\right]\right\} \ldots (9)$$

We obtain the maximum partial likelihood estimator by differentiating the right hand side of with respect to β, setting the derivative equal to zero and solving for the unknown parameter. The derivative of (9) with respect to β is

$$\frac{\partial L_p(\beta)}{\partial \beta} = \sum_{i=1}^{m}\left\{x_{(i)} - \frac{\sum_{j\epsilon R(t_{(i)})} x_j e^{x_i\beta}}{\sum_{j\epsilon R(t_{(i)})} e^{x_i\beta}}\right\} \quad \ldots (10)$$

$$= \sum_{i=1}^{m}\left\{x_{(i)} - \sum_{j\epsilon R(t_{(i)})} w_{ij}(\beta)\, x_j\right\}$$

$$= \sum_{i=1}^{m}\left\{x_{(i)} - \overline{x_{w_i}}\right\}$$

All software packages provide the maximum partial likelihood estimator, which we denote as $\hat{\beta}$. After fitting the proportional hazards regression model, the next step is to assess the overall significance of the model. The log partial likelihood ratio test is used to assess the overall significance of the model and it is denoted as

$$G = 2\left\{L_p(\widehat{\beta}) - L_p(0)\right\} \quad \ldots (11)$$

For this test the null hypothesis is

H_0: The covariates in the model are not significantly related to survival time.

H_1: The covariate in the model are significantly related to survival time.

We compare G with chi-square distribution with k degrees of freedom, where k is the number of covariates in the model. If the calculated value is less than G value then accept the null hypothesis. Otherwise reject the null hypothesis. To assess the significance of the covariates we use the p-values. If the p-values are less than 0.005 then the covariate is significant. Otherwise, we remove the covariate from the model. Then we fit a revised model and assess the significance of the model.

Kaplan-Meier Curve:

Assume that we have a sample of n independent observation denoted (t_i,c_i), i=1,2,…,n of the underlying survival time variable T and the censoring indicator variable C. Assume that, among the n observation, there are m≤n recorded times of failure and n-m censored values. We denote the rank-ordered survival times as $t_{(1)}<t_{(2)}<\ldots<t_{(m)}$. In this text, when quantities are placed in rank order, we use the same variable notation but place subscripts in parentheses. Let the number at risk of dying at $t_{(i)}$ be denoted n_i and the observed number of deaths be denoted d_i. The Kaplan-Meier estimator of the survival function at time t is

$$\hat{S} = \prod_{t_{(i)}\le t}\frac{n_i - d_i}{n_i}$$

In the analysis of survival time, the sample mean is not as important measure of central tendency as in other settings (the exception is in fully parametric modeling of survival times when the estimator of the mean, or a function of it provides an estimator of a parameter vital to the analysis and interpretation of the data). This is due to the fact that censored survival time data are most often skewed to the right, a setting whether the median usually provides a more intuitive measure of central tendency. We have to calculate both the mean and median survival time for the variables. When comparing group of subjects, we should begin with the graphical display of the data in each group. In studies of survival time, we should graph the Kaplan-Meier estimator of the survival function for each of the groups. The statistical question is whether the observed difference shown by the Kaplan-Meier graph is statistically significant or not.

Results and Discussion:

In this work fitting of cox regression model for the heart failure dataset with all the covariates to find which variables cause death at most or which variables were significant in the model were discussed. The overall model significance also evaluated. As described in Section 2, a cox regression model was fitted considering all the eleven predictors as significant for diagnosing cardiac illness. The overall significance of the fitted Model may be assessed by testing

H0: βj = 0, j = 1, 2, …, k againstH1: βj ≠ 0, at least for some j.

Table 1 : Omnibus Tests of Model Coefficients

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block
	Chi-square	Df	Sig.	Chi-square	df	Sig.	Chi-square
937.572	87.670	11	,000	81.171	11	,000	81.171

Table 1 shows that value of the -2Log(Likelihood) test statistic computed for the fitted model is calculated as 937.572. Under H_0, the sampling distribution of the test statistic χ^2_{11} at 5% level of significance, the value of the test statistic is approximately, 19.6751. Hence, the sample information provides larger evidence for rejecting H_0. Therefore, it can be inferred as the fitted model has overall significance.

Table 2: Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
Age	.046	.009	24.519	1	.000	1.047
Anaemia	-.455	.217	4.401	1	.036	.634
creatinine phosphokinase	.000	.000	4.973	1	.026	1.000
Diabetes	-.137	.223	.377	1	.539	.872
ejection fraction	-.049	.011	21.592	1	.000	.952
high_blood_pressure	-.472	.216	4.755	1	.029	.624
Platelets	.000	.000	.210	1	.647	1.000
serum_creatinine	.312	.070	20.029	1	.000	1.366
serum_sodium	-.044	.023	3.652	1	.05	.957
Gender	.240	.252	.909	1	.340	1.271
Smoking	-.119	.251	.223	1	.637	.888

In the table 2 the Exp(B) referred the hazard ratio. If the hazard ratio is greater than one then the predictor has increased risk of death. Among the significant variables age and serum creatinine had hazard value more than one. Ejection fraction and serum sodium has almost one. Thus age, serum creatinine, ejection fraction and serum sodium increase the risk of death due to heart disease.

Conclusion:

The Kaplan-Meier analysis was performed for all the eleven independent variable. Among them age, ejection fraction, serum sodium and serum creatinine were significant. The survival time of the patients were differ groups to group in the four variables. Thus we conclude that these four variables play an immense role to determine the survival time of the heart disease patients. The Cox regression model was fitted to the heart failure dataset with eleven predictor variables. The Cox regression model was fitted to identify which factor affect the survival time of heart disease patients. The significant predictors age, ejection fraction, serum creatinine and serum sodium were the factors which determines the survival time of the heart diseased patient. In multivariate Cox analysis, the p-value of age is zero with hazard ratio is 1.047 indicating the strong relationship between age and decreased risk of death. It indicates the younger people who had heart disease has longer life time when compared to elder people.

The p-value of ejection fraction is zero with hazard ratio 0.952 indicating the strong relationship between ejection fractions increased risk of death. The co-efficient value is negative and it indicates that increased risk of death. It indicates that if the ejection fraction level is low then the risk of death is high. The p-value of serum creatinine is zero with hazard ratio 1.366 indicating the strong relationship between serum creatinine and decreased risk of death. It indicates that who has normal serum creatinine level has longer life time than the patient who has up normal serum creatinine level.

The p-value of serum sodium is 0.05 with hazard ratio with 0.957 indicating the strong relationship between serum sodium and decreased the risk of death. It indicates that who has normal serum sodium level has longer life time than who has normal up serum sodium level. From Cox regression model age, ejection fraction, serum creatinine and serum sodium were the factors which decides the life span of a heart diseased patient among the eleven factors. All the result were given the same result that age, ejection fraction, serum creatinine and serum sodium has the great impact on the survival time of the heart disease patients.