تقدير نموذج الانحدار المعلمي بالاعتماد على توزيع وقت البقاء على قيد الحياة (وقت الحدث) مع التطبيق

Time Estimate Parametric Regression Models Depend on (Time-To-Event) Survival Distributions with Application

دانا طه محمد صالح
Dana Taha Mohammed Salih
dana.taha.@uoh.edu.iq
جامعة حلبجة
University of halabja

آ.م.د.سميرة محمد صالح محمد
Asst. Prof. Dr. Samira Muhammad Salh
samira.muhamad@univsul.edu.iq
جامعة السليمانية
University of Sulaimani

آ.م.د.محمد محمود فقى
Asst. Prof. Dr. Mohammad Mahmood Faqe
mohammad.faqe@univsul.edu.iq
جامعة السليمانية
University of Sulaimani

الملخص

هدفت الدراسة الى المقارنة بين نماذج الانحدار المعلمية المقدرة وفقًا لتوزيعات وقت البقاء واختيار أفضل الانموذج الملائم لتوزيع وقت البقاء على قيد الحياة وتقدير معلماتها .واستعمال نماذج الانحدار المعلمى لبيانات و مدى تحديد العوامل التي تؤثر في وقت بقاء المرضى .وتم تطبيق الدراسة على عينة بحجم (١٢٠) مرضا، المصابون بسرطان بروستات/ مستشفى هيوا فى محافظة السليمانية لمدة من 1 يناير 2019 حتى 1 نوفمبر 2021.

وتم تحديد افضل انموذج بالاعتماد على كل من المقايس (AIC,BIC) وباستعمال البرامج (Mat-lab, Stata 15.1, Easy Fit 5.6). وتبين ان افضل انموذج هو (Weibull- AFT) والعوامل (Age, PSA, Stage, metastasis)هي التي تؤثر في مدة البقاء المريض.

الكلمات الافتتاحية : تحليل البقاء على قيد الحياة، نماذج الانحدار المعلمية، وقت الفشل المعجل (AFT)، MLE، سرطان بروستات.

Abstract:

The aim of this study is to compare between the parametric regression models estimated according to the distributions of survival time and select the best appropriate model for the distribution of survival time and estimate its parameters. Use parametric regression models for data and determine the factors that affect the survival time of patients. The study was applied to a sample size of (120) patients with prostate cancer / Hiwa Hospital in Sulaymaniyah Governorate for a period from January 1, 2019 to November 1, 2021.

The best model was determined based on each of the criteria AIC and BIC and using the applications (Mat-lab, Stata 15.1, Easy Fit 5.6). The result show that the best model is (Weibull-AFT) and the factors (Age, PSA, Stage, metastasis) are that affect the patient's survival time.

Keywords: Survival Analysis, parametric regression models, Accelerated Failure Time (AFT), Maximum likelihood estimation (MLE), prostate Cancer.

1: Introduction

Survival analysis is a statistical tools analysis that focuses on the influence of predictors on the time until an event happens, rather than the chance of an event occurring. It is used to examine data that contains information about the time remaining until an event occurs. As the name suggests, this approach was usually used in medical research to evaluate the effect of drugs or medical therapy on the time before death. In the engineering field is called reliability analysis, it's also called duration analysis in the economics field, and it's called event history analysis in the sociology field. This method was used to measure the survival rate of patients in the medical field. In survival analysis models, the process and paradigms that can be used to deal with data classification can be used. Survival analysis may be studied using a variety of methods, including(Life tables, Kaplan-Meier analysis, Survivor and hazard function rates, Cox proportional hazards regression analysis, parametric survival analytic models, Survival trees, Survival random forest) (Abbas, Subramanian et al. 2019).

In scientific and organic studies, the analysis of event time data or (survival statistics) aimed to describe the hazard (risk) function of event times in population. Survival analysis is used in a lot of different fields, like biology, epidemiology, medicine, and public health. Survival data is often examined by simulating event timing data, such as the amount of time till death. Survival time or failure time is the term used to describe the amount of time before a particular occurrence occurs(Zhao 2008).

Survival analysis is a method for analyzing time-to-event data, data where the outcome variable is time elapsed from a time origin until the occurrence of a chosen event of interest. This type of data is common in medical studies where often the time origin

corresponds to entry into the study and the event of interest to death, thus the name survival analysis (Christiansson 2020). At the begging survival used only to investigate mortality and morbidity in vital statistics. . The first mathematical examination of human survival processes dates all the way back to the seventeenth century, when John Graunt, an English statistician, produced the first life table in 1662(Liu 2012).

2: Literature review

Jiezhi Qi (2009), Compared Proportional Hazards and Accelerated Failure Time Models, In the study of some survival data, the AFT model was considered as an alternative to the PH model(Qi 2009).

Yiu Ming Chan (2013), this investigation gave a correlation of survival between White and African American men at the four key stages of cancer for patients under a similar treatment. Moreover, to understand the hazard factors (age, tumor estimate, and tumor size connected with survival time, a system of accelerated disappointment time was made. Finally, the results of parametric survival examination and the system of accelerated disappointment time model are looked at among white men experiencing a similar treatment(Chan 2013) .

Minh Hoang Pham (2014), this study observed the cancer patients' survival time using the Weibull probability distribution. In this study, the researcher made use of the parametric survival method for analysis of the survival time cancer patients. The results of the study prove consistency of the parametric method in relation with the theoretical approach more than the semi-parametric approach(Pham 2014)

Montaseri, Charati, and Espahbodi (2016), the purpose of this research was to examine the performance of several parametric models in a survival analysis of patients undergoing hemodialysis. Parametric models were compared using the Akaike information criterion (AIC). It was shown that the mean serum albumin and attendance at a clinic were the most significant predictors of hemodialysis patient. According to the findings of the parametric models evaluated, the Weibull model had the best performance(Montaseri, Charati et al. 2016)

Emmert-Streib and Dehmer (2019), in this research, reviewed the theoretical foundations of survival analysis, including estimators for survival and hazard functions, have been discussed in this study. They used the Cox Proportional Hazard Model, and also stratified Cox models was used for when the PH assumption doesn't hold. (Emmert-Streib and Dehmer 2019).

Salinas-Escudero...etal (2020), the purpose of this investigation is to use survival analysis to determine the risk variables related with COVID-19 mortality in the Mexican population. They used this analysis to make Kaplan-Meier curves and a Cox proportional hazard model.Concluded that men had a higher risk of dying at any time during follow-up than women with the patients over the age of 65, adults with chronic renal disease. (Salinas-Escudero, Carrillo-Vega et al. 2020).

3: Methodology

This section discussed the some basic definitions of survival analysis, including the nature of data (typed of censoring), survival function and the hazard function, as well as several tests and techniques for analyzing survival data.

3.1: Survival Analysis

Survival analysis is a collection of methods for studying data in which the outcome variable is the time until an event of interest occurs. The occurrence might be death, divorce, the onset of a sickness, marriage, and so on. Years, months, weeks, Days, and so on can be used to calculate the time to event or survival time(Ekman 2017).

3.2: The nature of survival data (Censoring data)

In Survival analysis there are several types of data was used to analysis, below definitions of these different types of censoring:

3.2.1: Type I Censoring

In this type the study comes to an end at a certain time point or, if the participants are tested at multiple times during the research, when a specific amount of time has passed from the beginning of the experiment(Ekman 2017).

3.2.2: Type II Censoring

The censoring of failure time data sets it apart from other data types. Assume we study the mortality rates of individuals with a certain condition. It is normal that some

patients be still alive at the conclusion of the trial. So their failure times are known to be larger than the duration from patient enrollment to study completion. As a result of this censoring, survival analysis requires statistical techniques other than simple linear regression. There are three kinds of censoring: right, left, and interval(Dey, Mukherjee et al. 2020).

3.2.2.1: Right Censoring

If failure happens after the documented follow-up period, a subject is right censored(Stevenson and EpiCentre 2009)

3.2.2.2: Left censoring

Left censoring occurs when the event is already past. This is an uncommon occurrence. Assume that some individuals in the stroke clinical trial experienced a stroke before the research began. These subjects are left-censored observations, where the "failure" (stroke) happened prior to a certain time. A subject is left censored it is known that the failure occurs some time before the recorded follow-up period(Dey, Mukherjee et al. 2020).

3.2.2.3: Interval censoring

It is described as interval censored when the event happens between two times, but the actual moment of failure is unknown. In other words, I can tell that the event happened between the dates A and B(Alhasawi 2015).

3.3: Failure Time

Usually the failure time of a survival depends on time, with the rate varying over the life cycle of the survival. It is interested in the effect of a risk factor or therapy on the time required to develop a disease or other occurrence(Vittinghoff, Glidden et al. 2006).

3.4: Function Related to Survival Analysis.

3.4.1: Cumulative distribution

The cumulative distribution is defined as follows:

$$F(t) = pr(T \leq t) = \int_0^t f(u)du \qquad (1)$$

The time interval is expected to be between 0 and **t.** (Abbas, Subramanian et al. 2019).

3. 4.2: Survival function

Survival probability is produced by the survival function, **S (t)**, approximately to time **t.** for survival analysis, survival function has an important role to play. **T** is a random variable that refers to the survival time, whereas **S(t)** refers to the survival function and **(T)** is a non-negative random variable referring to the time when an event occurs, The definition of the survival function appears to be as follows(Emmert-Streib and Dehmer 2019) :

$$S(t) = pr(T > t) = 1 - F(t) = \int_t^{\infty} f(u)du \qquad (2)$$

3.4.3: Hazard Function

We define the hazard

function and the relationship between it and the survival function in this section of the paper. The following is the definition of the hazard function(Lee and Wang 2003):

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t | T \ge t)}{\Delta t}$$

$$h(t) = \lim_{\Delta t \to 0} \frac{[F(t + \Delta t) - F(t)]/\Delta t}{S(t)}$$

$$h(t) = \frac{\frac{\partial F(t)}{\partial t}}{S(t)}$$

$$h(t) = \frac{f(t)}{S(t)} \qquad (3)$$

3.5: Models of Survival Analysis

3.5.1: Parametric Models

Parametric approaches are based on the assumption that the fundamental distribution of survival times follows well-known probability distributions. Including(exponential ,Weibull, and lognormal distributions), with Accelerated Failure Time **(AFT)** model (Wang, Li et al. 2019).

3.5.1.1: Accelerated Failure Time (AFT)

It is another popular regression model, often, used to analyze survival data, also, AFT model relate the lifetime distribution to the explanatory variable (stress, covariate).

This distribution can be defined by the survival, cumulative distribution, or probability density functions(Bogaerts, Komárek et al. 2017).

Regarding T_i as a random variable representing the (possibly unobserved) survival time of the i[th] unit, since T_i must be non-negative value, and it should be considered modeling its logarithm using a customary linear model:

$$log\, T_i = x_i{}' \beta + \varepsilon_i \qquad (4)$$

Where:

ε_i is advisable error term and x_i is covariate factor, T_i is survival time.

The distribution of survival time to be specified (exponential, Weibull, log-normal and gamma AFT model) (Cleves, Gould et al. 2008).

3.5.1.1.1: Weibull distribution(Lee and Wang 2003)

The probability density function (**p.d.f**) and cumulative distribution functions (**C.D.F**) are, respectively:

$$f(t) = \frac{\theta}{\delta}(\frac{t}{\delta})^{\theta-1} e^{-(\frac{t}{\delta})^{\theta}} \qquad t \geq 0, \quad \theta, \delta > 0 \qquad (5)$$

And

$$F(t) = 1 - e^{-(\frac{t}{\delta})^{\theta}} \qquad (6)$$

The survival function is, therefore,

$$S(t) = e^{-(\frac{t}{\delta})^{\theta}} \qquad (7)$$

And the hazard function, the ratio of (5) to (7), is

$$h(t) = \frac{f(t)}{s(t)}$$

$$h(t) = \frac{\frac{\theta}{\delta}\left(\frac{t}{\delta}\right)^{\theta-1} e^{-\left(\frac{t}{\delta}\right)^{\theta}}}{e^{-\left(\frac{t}{\delta}\right)^{\theta}}}$$

$$h(t) = \frac{\theta}{\delta}(\frac{t}{\delta})^{\theta-1} \qquad (8)$$

The mean of the Weibull distribution is

$$\mu = \delta\Gamma\left(1 + \frac{1}{\theta}\right) \qquad (9)$$

And the variance is

$$\sigma^2 = \delta^2\left[\Gamma\left(1 + \frac{2}{\theta}\right) - \Gamma^2\left(1 + \frac{1}{\theta}\right)\right] \qquad (10)$$

Where $\Gamma(\boldsymbol{\theta}) is\ the\ well-known\ gamma\ function\ defined\ as$

$$\Gamma(\theta) = \int_0^{\infty} x^{\theta-1} e^{x}\, dx$$
$$= (\theta - 1)! \qquad (11)$$

Value of $\Gamma(\boldsymbol{\theta})$ can be found in Abramowitz and Stegun (1964). The coefficient of variation is then

$$C.V = \frac{1}{\delta}\left[\frac{\Gamma\left(1+\frac{2}{\theta}\right)}{\Gamma\left(1+\frac{1}{\theta}\right)} - 1\right]^{\frac{1}{2}}$$

(12)

3.6: Maximum likelihood estimation (MLE)

Now we are using MLE to estimate to parameters of Weibull distribution

$$f(t) = \frac{\theta}{\delta}\left(\frac{t}{\delta}\right)^{\theta-1} e^{-\left(\frac{t}{\delta}\right)^{\theta}} \qquad t \geq 0, \quad \theta, \beta > 0$$

$$L(\theta, \delta, t_1, t_2, \dots, t_n) = \left(\frac{\theta}{\delta}\right)^{n} \prod_{i=1}^{n} \left(\frac{t_i}{\delta}\right)^{\theta-1} e^{-\sum_{i=1}^{n}\left(\frac{t_i}{\delta}\right)^{\theta}}$$

$$lnL(\theta, \delta, t_1, t_2, \dots, t_n) = nln\left(\frac{\theta}{\delta}\right) + (\theta - 1) ln \prod_{i=1}^{n} \left(\frac{t_i}{\delta}\right) - \sum_{i=1}^{n} \left(\frac{t_i}{\delta}\right)^{\theta}$$

$$lnL(\theta, \delta, t_1, t_2, \dots, t_n) = \; nln\,(v) - n\ln(\delta) + \theta \sum_{i=1}^{n} ln\left(\frac{t_i}{\delta}\right) - \sum_{i=1}^{n} \ln\left(\frac{t_i}{\delta}\right) - \sum_{i=1}^{n} \left(\frac{t_i}{\delta}\right)^{\theta}$$

$$lnL = \; nln\,(\theta) - n\ln(\delta) + \; \theta \sum_{i=1}^{n} (\ln t_i - ln\delta) - \sum_{i=1}^{n} (\ln t_i - ln\delta) - \sum_{i=1}^{n} \left(\frac{t_i}{\delta}\right)^{\theta}$$

$$lnL = \; nln\,(\theta) - n\,ln(\delta) + \; \theta \sum_{i=1}^{n} lnt_i - \theta \sum_{i=1}^{n} ln\delta - \sum_{i=1}^{n} \ln t_i + \sum_{i=1}^{n} \ln \delta - \sum_{i=1}^{n} t_i^{\theta}\; \delta^{-\theta}$$

Now we get derivative by $\boldsymbol{\delta}$:-

$$\frac{d\; lnL}{d\; \delta} = \frac{-n}{\delta} - \theta \sum_{i=1}^{n} \frac{1}{\delta} + \sum_{i=1}^{n} \frac{1}{\delta} - \sum_{i=1}^{n} t_i^{\theta}\; (-a\delta^{-(\theta+1)})$$

$$\frac{d\; lnL}{d\; \delta} = \frac{-n}{\delta} - \frac{\theta n}{\delta} + \frac{n}{\delta} + \theta \sum_{i=1}^{n} \frac{t_i^{\theta}}{\delta^{\theta+1}}$$

$$\frac{d\; lnL}{d\; \delta} = -\frac{\theta n}{\delta} + \frac{\theta}{\delta} \sum_{i=1}^{n} \left(\frac{t_i}{\delta}\right)^{\theta}$$

$$\frac{d\; lnL}{d\; \delta} = 0$$

$$-\frac{\theta n}{\hat{\delta}}+\frac{\theta}{\hat{\delta}}\sum_{i=1}^{n}(\frac{t_i}{\delta})^{\theta}=0$$

$$\frac{\theta}{\hat{\delta}}\ (\sum_{i=1}^{n}(\frac{t_i}{\delta})^{\theta}-n)=0$$

$$\hat{\delta}^{\theta}=\sum_{i=1}^{n}\frac{{t_i}^{\theta}}{n}$$

$$\hat{\delta}=\left(\sum_{i=1}^{n}\frac{{t_i}^{\theta}}{n}\right)^{\frac{1}{\theta}}$$

(13)

And we get derivative with respect $\boldsymbol{\theta}$:-

$$\frac{d\ lnL}{d\ \theta}=\frac{n}{\theta}+\sum_{i=1}^{n}\ln(\frac{t_i}{\delta})-\ \delta\sum_{i=1}^{n}(\frac{t_i}{\delta})^{\theta}\ln(\frac{t_i}{\delta})$$

$$\frac{d\ lnL}{d\ \theta}=0$$

$$\frac{n}{\hat{\theta}}+\sum_{i=1}^{n}\ln(\frac{t_i}{\delta})-\ \delta\sum_{i=1}^{n}(\frac{t_i}{\delta})^{\hat{\theta}}\ln\left(\frac{t_i}{\delta}\right)=0$$

We can now assume that $\boldsymbol{h}(\widehat{\boldsymbol{\theta}})$ is the function of its partial derivative $\frac{\boldsymbol{d\ lnL}}{\boldsymbol{d\ \theta}}$, where

$$h(\hat{\theta})=\frac{n}{\hat{\theta}}+$$

$\sum_{i=1}^{n}\ln(\frac{t_i}{\delta})-\ \delta\sum_{i=1}^{n}(\frac{t_i}{\delta})^{\hat{\theta}}$ (14)

Because of the difficulty of solving the equation (14) in the usual methods, we will use iterative methods, including Newton-Raphson-method, to obtain an estimate of $(\boldsymbol{\theta},\boldsymbol{\delta})$ the steps of the method depend on the assumption of an (initial value) of the required root $(\widehat{\boldsymbol{\theta}})$ using the OLS method and be $(\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}=\widehat{\boldsymbol{\theta}})$ and then determine the roots approximate to $(\widehat{\boldsymbol{\theta}})$ as in the following equation:

$$\hat{\theta}_{i+1}\ =\hat{\theta}_{\iota}-\frac{h(\hat{\theta}_i)}{h'(\hat{\theta}_i)}$$

(15)

Where $\boldsymbol{h}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}})$ represent equation (14) and $h'(\widehat{\theta}_{\iota})=\frac{dh(\hat{\theta}_i)}{d(\hat{\theta}_i)}$

At first we impose an initial value which is $(\widehat{\boldsymbol{\theta}}_{\boldsymbol{0}})$ where $i=\ 0$ and then apply an equation (15) to get a new value to $\widehat{\boldsymbol{\theta}}$ and be $\widehat{\boldsymbol{\theta}}_{\boldsymbol{1}}$ and then assume that $\hat{\theta}_1$ is the initial value and apply the equation (15) again to get a new value which is $\widehat{\boldsymbol{\theta}}_{\boldsymbol{2}}$ and so until we reach the

stage $(\mathbf{i+1})$ then $\widehat{\boldsymbol{\theta}}_{\boldsymbol{i+1}}$ approach the required degree of accuracy specified by the researcher and thus we get the estimation of $\widehat{\boldsymbol{\theta}}$ which represents the greatest value

$$h'(\hat{\theta}) = -\frac{n}{\widehat{\theta^2}} - \sum_{i=1}^{n} (\frac{t_i}{\delta})^{\hat{\theta}} [\ln(\frac{t_i}{\delta}) \ln(\frac{t_i}{\delta})$$

$$h'(\hat{\theta}_i) = -\frac{n}{\widehat{\theta^2}_i} - \sum_{i=1}^{n} (\frac{t_i}{\delta})^{\hat{\theta}} [\ln\left(\frac{t_i}{\delta}\right)]^2$$

(16)

We stop when:-

$$|\hat{\theta}_{i+1} - \hat{\theta}_i| \leq e$$

Equal to a very small value and then in order to get an initial value to apply the Newton-Ravson method. I will use the method of OLS to obtain this value, but the way to find this value depends on the function of the cumulative distribution function of Weibull and its formula is as follows:

$$F(t) = 1 - e^{-(\frac{t}{\delta})^{\theta}}$$

$$e^{-(\frac{t}{\delta})^{\theta}} = 1 - F(t)$$

And take $(\boldsymbol{ln})$ to the both side of the equation:-

$$-\left(\frac{t}{\delta}\right)^{\theta} = ln[1 - F(t)]$$

$$\left(\frac{t}{\delta}\right)^{\theta} = -ln[1 - F(t)$$

And again $\boldsymbol{ln}$ to the both side

$$\theta \ln(\frac{t}{\delta}) = \ln[-ln(1 - F(t))]$$

$$\theta ln(t) - \theta \ln(\delta) = \ln[-ln(1 - F(t))]$$

$$ln(t) - \ln(\delta) = \frac{1}{\theta} \ln[-ln(1 - F(t))]$$

$$ln(t) = \ln(\delta) + \frac{1}{\theta} \ln[-ln(1 - F(t))]$$

(17)

And by comparing the equation (17) with the equation of simple linear regression:-

$$Y_t = B_0 + B_1 X_t + E_i \qquad i=1,2,3,\ldots,n$$

(18)

Where $\boldsymbol{E_i}$ represents random error

$Y_t = ln(t)$

$X_t = \ln[-ln(1 - F(t))]$

$$B_0 = \ln(\delta)$$

$$B_1 = \frac{1}{a}$$

Cumulative distribution $\boldsymbol{F(t)}$ values can be obtained from empirical distribution and according to the following formula:-

$$F(t) = \frac{J-0.5}{N} \quad , \text{J=1,2,...,N} \quad (19)$$

$$\widehat{B}_0 = \bar{X} - \hat{B}_1 \bar{X} \quad (20)$$

$$\hat{B}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{Y}\bar{X}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} \quad (21)$$

And then we can get Weibull distribution from the following relationships:-

$$\hat{B}_0 = \ln(\hat{\delta}) \quad , \quad \hat{\delta} = e^{\hat{B}_0} \quad (22)$$

$$\hat{B}_1 = \frac{1}{\hat{\theta}} \quad , \hat{\theta} = \frac{1}{\hat{B}_1} \quad (23)$$

3.7: Measures of the Model Selection

Two criteria were used in this research to determine the best model. Which are (**AIC**) and (**BIC**) criteria, we prefer lowest value to choose the best model. The two criteria we use are the following:

3.7.1: Akaikke's Infirmation Criterion (AIC)

Comparing the quality of various statistical models is done using the Akaike Information Criterion (AIC). In comparison to other models, the model with a lower AIC fits the data better.

Akaike's Information Criterion calculating as follows:

$$\text{AIC} = -2(\text{log-likelihood}) + 2K \quad \text{or} \quad \text{AIC} = 2K - 2\ln(L) \quad (24)$$

Where:

K the number of parameters in the model (the model's total number of variables, plus the intercept).

Log-likelihood is a measure of model fit, this is usually found in the statistical output(Moore 2016).

3.7.2: The Bayesian Information Criterion (BIC)

The Bayesian information criterion (BIC) is a well-known and commonly used tool for selecting statistical models. Bayesian information criterion calculate as follows:

$$BIC=-2*\ln L+2*\ln N*k \quad (25)$$

Where: **L**=is the value of the likelihood

N=the number of observations, or equivalently, the sample size.

K= number of model parameters.

the model with the lowest BIC is chosen as the best model(Ibrahim, Chen et al. 2001).

4: Description and Analysis of data:

4.1: Data description:

The data for this paper of prostate cancer have been collected from Hiwa Hospital. The data consisted of 120 cases which are collected during 3 years period; beginning from 1th January 2019 through 1th November 2021 on all prostate cancer patients. Out of those patients there are 120 patients are survived or still alive. The survival time are measured in day.

Table 1. The explanatory variables measured for these data at diagnosis:

Name variables	Description	Percentage (%)
Age	<60=(1)	%7.5
	60-69=(2)	%32.5
	70-80=(3)	%43.3
	>80=(4)	%16.7
Smoker	Yes=(1)	%27.5
	No=(2)	%22.5
Blood group	A+=(1)	%24.1
	A-=(2)	%1.7
	B+=(3)	%14.2
	B+=(4)	%3.3
	O+=(5)	%47.5
	O=(6)	%2.5

	AB+=(7) AB=(8)	%6.7 %0
Occupation	Employee=(1) No employee=(2) retired=(3)	%5.8 %42.5 %51.7
stage	Lowe-risk=(1) intermediate-risk=(2) high-risk=(3)	%1.7 %17.5 %80.8
Volume	≤40 =(1) 41-60=(2) >60 =(3)	%38.3 %27.5 %34.2
BMI	Underweight=(1) normal weight=(2) overweight=(3)	%1.7 %42.2 %45.2
PSA	<10 =(1) 10-20 =(2) 20> =(3)	%7.5 %20.8 %71.7
Metastasis	Yes=(1) No=(2)	%66.7 %33.3
Genetic	Yes=(1) No=(2)	%20 %80

And the **response variable** is survival (Time-To-Event)

4.2: Data Analysis:

First of all we test the data to know that if this data Weibull distribution or not, we use the goodness of fit, which is Kolmogorov-Smirnov, Anderson-Darling, Chi-Squared test, according to our hypothesis. Table below show the result of this test Hypothesis test:

H_0:The data distributed Weibull distribution

H_1: The data distributed Weibull distribution

Table 2: Test data for parametric regression Weibull distribution.

	$a = 0.01$			$a = 0.05$		
	Chi-Squared	Anderson-Darling	Kolmogorov-Smirnov	Chi-Squared	Anderson-Darling	Kolmogorov-Smirnov
Statistic	8.0823	0.5173	0.06426	8.0823	0.5173	0.06426
Critical Value	16.812	3.9074	0.14871	12.592	2.5018	0.12397

The table above show that the critical value of Chi-Squared, Anderson-Darling and Kolmogorov-Smirnov $(16.812,\ 3.9074,\ 0.14871)$ greater than their statistics $(8.0823,\ 0.5173,\ 0.06426)$ with $(a = 0.01)$ and $(12.592,\ 2.5018,\ 0.12397)$ greater than $(8.0823,\ 0.5173,\ 0.06426)$ therefore accept H_0 and the survival time follows the Weibull distribution.

Second step: **fitting the accelerated Failure time (AFT)** model with (Weibull, Log normal, exponential) distribution.

Table 3: The (BIC) and (AIC) tests, for comparing AFT Model.

Distribution	NO. parameter	AIC	BIC
Weibull	2	232.695	266.1449
Log normal	2	267.4484	300.8983
exponential	1	303.9187	334.5811

According to the **Table (3)** compared AFT models by statistics criterion Bayesian information criterion (BIC) and Akaike information criterion (AIC). The smaller BIC and AIC is the better, each of the BIC and the AIC are tools for choosing between two or more models.in the above table explained that the Weibull AFT model is better model according to AIC=232.695 and BIC=266.1449 compared with models.

Third step: Finding the initial value for Weibull distribution

To obtain the initial value by use equation (18) and the equation $Y_t = ln(t)$ and

$$X_t = \ln[-ln(1 - F(t))] .$$

By using ordinary least square (OLS) to estimate of values (B_0, B_1) we get the value of

$$B_0 = 5.962,\ B_1 = 0.72$$

Table 4: show that the transformation for the explanatory variable (X_t) and Response variable (Y_t)

J	t	$Y_t = ln(t)$	$F(t) = \frac{J-0.5}{N}$	1-$F(t)$	$ln(1-F(t))$	-$ln(1-F(t))$	$X_t = \ln(-ln(1-F(t))$	X_t^2	$X_t \times Y_t$
1	219	5.38907	0.00417	0.99583	-0.00418	0.00418	-5.47855	30.01453	-29.5243
2	221	5.39816	0.01250	0.98750	-0.01258	0.01258	-4.37574	19.14713	-23.621
3	224	5.41165	0.02083	0.97917	-0.02105	0.02105	-3.86069	14.90495	-20.8927
4	227	5.42495	0.02917	0.97083	-0.02960	0.02960	-3.51997	12.39015	-19.0956
5	227	5.42495	0.03750	0.96250	-0.03822	0.03822	-3.26436	10.65608	-17.709
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.		.	.	.	.	.	.	.	.
118	972	6.87936	0.97917	0.02083	-3.87120	3.87120	1.35356	1.83214	9.3117
119	164	5.09987	0.98750	0.01250	-4.38203	4.38203	1.47751	2.18304	7.5351
120	164	5.09987	0.99583	0.00417	-5.48064	5.48064	1.70122	2.89416	8.6760

The coefficient of ordinary least square (OLS) method confirmed by use Stata program and then we applied the equation (22) and (23) to obtain the initial values of two parameters and the results were as follows:

$\theta_0 = 1.389$

$\delta_0 = 388.39$

At the end to find the value of Estimate shape and scale parameter of Weibull distribution:-

To find the values of the shape and scale parameters of Weibull distribution, we have used the method of (Newton-Raphson-method) with error (0.00001) to apply this method we use equation (14, 15, 16) the results of which were as follows:

$\theta = 1.6717$

$\delta = = 517.4$

Table 5: The survival model according to AFT Weibull distribution when the survival time followed the Weibull distribution.

Variables	Coef .	Std. Err.	Z	P>|z|	[99% Conf. Interval]	

					Lower bound	Upper bound	Haz. Ratio
Blood	0.0295271	0.0243332	1.21	0.225	−0.0331511	0.0922053	1.029967
Occupation	0 .1183633	0.0819883	1.44	0.149	−0.0928246	0.3295513	1.125653
Genetic	0.2437352	0.112841	2.16	0.031	−0.0469239	0.5343944	1.276006
Smoker	0.1203131	0.1040558	1.16	0.248	−0.1477169	0.3883431	1.12785
Stage	−0.6486137	0.167121	−3.88	0.000	−1.079089	−0.2181386	0.52277
Metastasis	−0.2950418	0.1113833	−2.65	0.008	−0.5819462	0.0081374	0.7445
Age	−0.0165876	0.0061613	−2.69	0.007	−0.0324581	−0.0007171	0.983549
PSA	0.3031847	0.106703	2.84	0.004	0.0283361	0.5780333	1.354165
Volume	−0.0010022	0.0011729	−0.85	0.393	−0.0040234	0.002019	0.998998
BMI	0.0176229	0.0106489	1.65	0.098	−0.0098067	0.0450526	1.017779

From the above table we notice the following:

Age variable will be one of the highly significant factors in our study; because the p−value of the variable is (0.007) less than the level of significance $(a = 0.01)$. Stage is the variable is significance because the p−value of the variable is less than the level of significance $(a = 0.01)$.

The p−value of variable (Metastasis=0.008) also is significance it is less than the level of significance $(a = 0.01)$.last significance variable (PSA) the p−value of the variable is (0.004) less than the level of significance $(a = 0.01)$.

Above table show that the PSA has the highest risk in prostate cancer which (1.354165) and stage has the lowest risk with rate (0.52277)

The variables (Blood, volume, BMI, genetic, occupation and smoker) are not significance factors while the P−value are greater than level significance $(a = 0.01)$,meaning that this variables are not affecting this type of cancer .

We can write the Weibull AFT model as follows:

$$\log T_i = x_i \beta + \varepsilon_i$$

$$\log T_i = -0.0165876\text{Age} \quad -0.3031847\text{PSA} - 0.6486137\text{stage} - 0.2950418\text{metatastis}$$

Table 6: Show that survival rate, hazard rate probability density function and cumulative

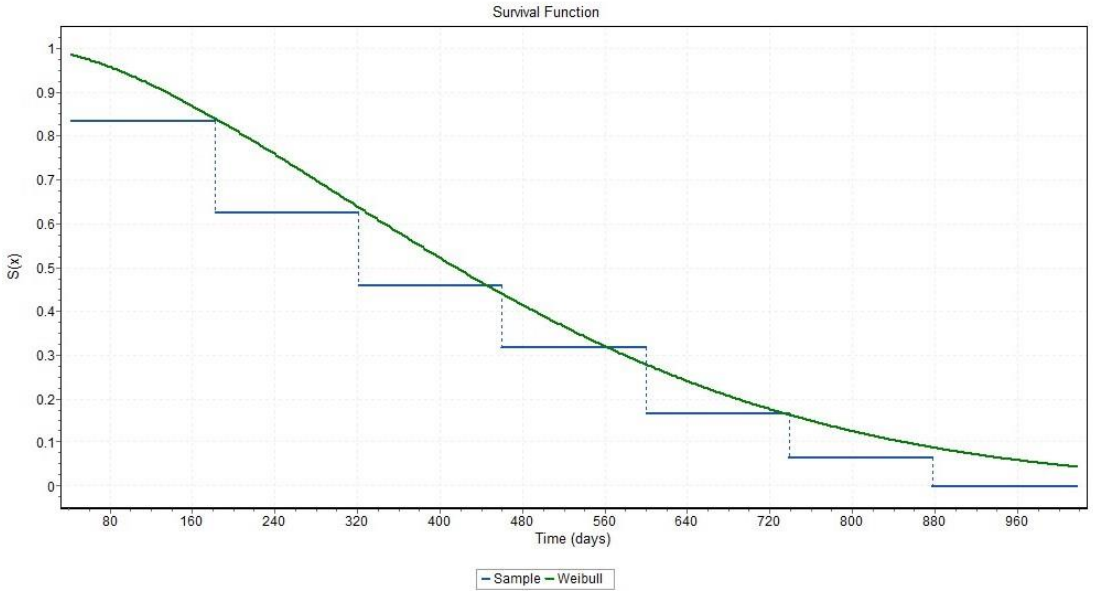
Time	f(t)	F(t)	h(t)	S(t)
42	0.00059	0.01491	0.00060	0.98509
43	0.00060	0.01551	0.00061	0.98449
45	0.00062	0.01672	0.00063	0.98328
55	0.00070	0.02331	0.00072	0.97669
57	0.00072	0.02473	0.00073	0.97527
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
1005	0.00024	0.95188	0.00505	0.04812
1018	0.00023	0.95495	0.00509	0.04505

of Weibull distribution.

The above table show that the value of survival function opposite with patients' stay in the hospital, and this means that the values of the survival function gradually decrease with the increase patients' stay in the hospital. If the patient's stay time is (42) days in the hospital, the probability of his survival is (0.98509), but if the patient's stay time is (1018) days in the hospital, the probability of his survival is (0.04505).

The values of the hazard function are positive and probabilistic values, and that the hazard function increases with the increase in the time patients stay in the hospital. Conversely, the longer the patient stays in hospital, the higher the risk of death. If the patient's stay time is (42) days in the hospital, the probability of death is (0.00060), but if the patient's stay time is (1018) days in the hospital, the probability of death is (0.00509).

The following two figure show the same results.



Figuer 1.Represent survival rate

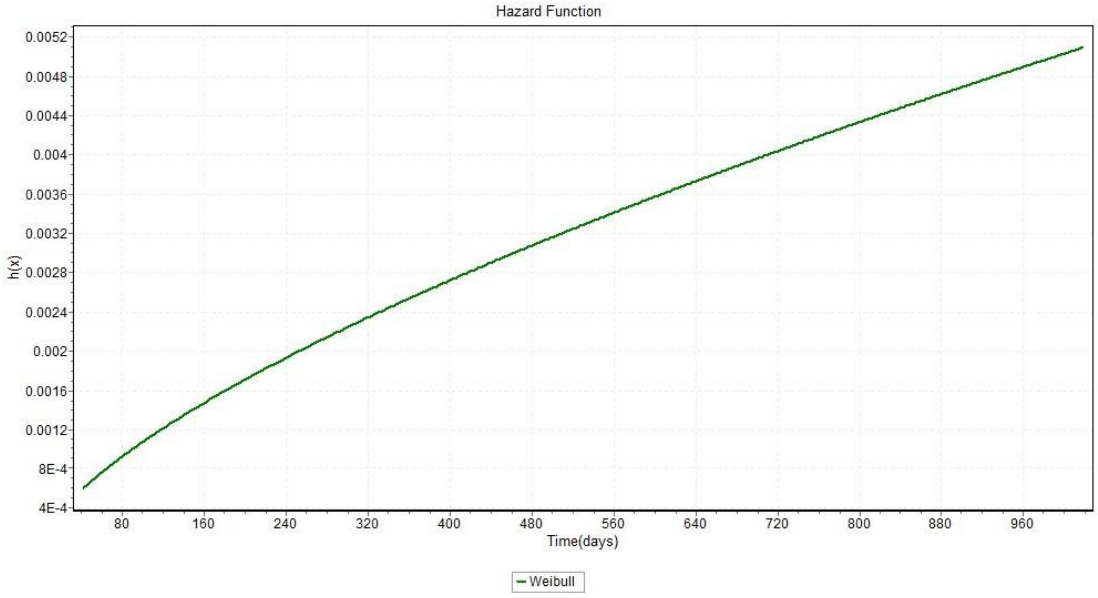


Figure 2 Represent hazard rate

5. Conclusion & Recommendation:

5.1: Conclusion:

During conducting the survival data and according to the results from the practical part the following conclusions have been shown:

1. Comparing the AFT models based on the AIC and BIC it is concluded that (Weibull AFT model) is the most suitable model for our data set that was used in this study.
2. According to the results of the Weibull AFT model of this study indicates the most popular factors that affected on the prostate cancer are (Age, PSA, Stage, metastasis) with level($a = 0.01$).
3. Determine the survival function and hazard function for each patients on prostate cancer. The results we get survival function gradually decrease with the increase patients' stay in the hospital. If the patient's stay time is (42) days in the hospital, the probability of his survival is (0.98509), but if the patient's stay time is (1018) days in the hospital, the probability of his survival is (0.04505). Hazard function increases with the increase in the time patients stay in the hospital which mean that the longer the patient stays in hospital, the higher the risk of death. If the patient's stay time is (42) days in the hospital, the probability of death is (0.00060), but if the patient's stay time is (1018) days in the hospital, the probability of death is (0.00509).

5.1: Recommendation

1. Using non-parametric models and comparing these models which we have achieved this study.
2. The variables we have achieved in this study should be considered by those who have a specialist in medicine in this field.
3. More studies should be done in this field because such studies are important and related to people's lives.
4. Data should be recorded in health places so that the researcher can conduct the research in detail.

7. Reference

Abbas, S. A., S. Subramanian, P. Ravi, S. Ramamoorthy and V. Munikrishnan (2019). An Introduction to Survival Analytics, Types, and Its Applications. Biomechanics, IntechOpen.

Alhasawi, E. (2015). Survival analysis approaches for prostate cancer, Laurentian University of Sudbury.

Bogaerts, K., A. Komárek and E. Lesaffre (2017). Survival analysis with interval-censored data: A practical approach with examples in R, SAS, and BUGS, Chapman and Hall/CRC.

Chan, Y. M. (2013). Statistical analysis and modeling of prostate cancer, University of South Florida.

Christiansson, A. (2020). Classication of survival data by comparison of survival functions: an application to prostate cancer registry data.

Cleves, M., W. Gould, W. W. Gould, R. Gutierrez and Y. Marchenko (2008). An introduction to survival analysis using Stata, Stata press.

Dey, T., A. Mukherjee and S. Chakraborty (2020). "A practical overview and reporting strategies for statistical analysis of survival studies." Chest **158**(1): S39-S48.

Ekman, A. (2017). Variable selection for the Cox proportional hazards model: A simulation study comparing the stepwise, lasso and bootstrap approach.

Emmert-Streib, F. and M. Dehmer (2019). "Introduction to survival analysis in practice." Machine Learning and Knowledge Extraction **1**(3): 1013-1038.

Ibrahim, J. G., M.-H. Chen, D. Sinha, J. Ibrahim and M. Chen (2001). Bayesian survival analysis, Springer.

Lee, E. T. and J. Wang (2003). Statistical methods for survival data analysis, John Wiley & Sons.

Liu, X. (2012). Survival analysis: models and applications, John Wiley & Sons.

Montaseri, M., J. Y. Charati and F. Espahbodi (2016). "Application of parametric models to a survival analysis of hemodialysis patients." Nephro-urology monthly **8**(6).

Moore, D. F. (2016). Applied survival analysis using R, Springer.

Pham, M. H. (2014). "Survival Analysis-Breast Cancer." Undergraduate Journal of Mathematical Modeling: One+ Two **6**(1): 4.

Qi, J. (2009). Comparison of proportional hazards and accelerated failure time models.

Salinas-Escudero, G., M. F. Carrillo-Vega, V. Granados-García, S. Martínez-Valverde, F. Toledano-Toledano and J. Garduño-Espinosa (2020). "A survival analysis of COVID-19 in the Mexican population." BMC public health **20**(1): 1-8.

Stevenson, M. and I. EpiCentre (2009). "An introduction to survival analysis." EpiCentre, IVABS, Massey University.

Vittinghoff, E., D. V. Glidden, S. C. Shiboski and C. E. McCulloch (2006). "Regression methods in biostatistics: linear, logistic, survival, and repeated measures models."

Wang, P., Y. Li and C. K. Reddy (2019). "Machine learning for survival analysis: A survey." ACM Computing Surveys (CSUR) **51**(6): 1-36.

Zhao, G. (2008). Nonparametric and parametric survival analysis of censored data with possible violation of method assumptions, The University of North Carolina at Greensboro.