EUROPEAN ACADEMIC RESEARCH
Vol. VII, Issue 1/ April 2019

Impact Factor: 3.4546 (UIF)
DRJI Value: 5.9 (B+)

ISSN 2286-4822
www.euacademic.org

Some Importance Models inspect to Survival Data Inference

ABU ELGASIM ABBAS ABOW MOHAMMED[1]
College of Business and Economics, Qassim University
Kingdom of Saudi Arabia
College of Economics and Political Science
Omdurman Islamic University, Sudan

Abstract

This paper used some models inspect to survival data analysis and focused on the Cox regression model characteristics. This paper also used liver cancer data from Khartoum State Health Ministry (Military Medical Hospital). The Cox regression model was estimated a number of descriptive variables such as sex, housing status, and 'quantitative variable such as age. The paper found that the age variable having a significant effect on the time of event, while the other variables have no significant effect

Keywords: Survival, baseline hazard, Cox regression, liver cancer

1. INTRODUCTION

There are many developments in the use of statistics in fields such as social sciences, economic, medical and other large-scale due to changes that occur in this area. This paper followed the most important models used in survival analysis data; it's called Cox's proportional hazards or Cox regression model. It is useful to begin by defining the concept of survival analysis.

[1] Corresponding author: gasimtas@gmail.com

Survival analysis is a branch of statistics for analyzing the expected duration of time until one or more events happen, such as death in biological organisms and failure in mechanical systems [1]. This topic is called reliability theory or reliability analysis in engineering, duration analysis or duration modeling in economics, and event history analysis in sociology. Survival analysis attempts to answer questions such as: what is the proportion of a population which will survive past a certain time? Of those that survive, at what rate will they die or fail? Can multiple causes of death or failure be taken into account? How do particular circumstances or characteristics increase or decrease the probability of survival? To answer such question, it is necessary to define "lifetime". In the case of biological survival, death is unambiguous, but for mechanical reliability, failure may not be well-defined, for there may well be mechanical systems in which failure is partial, a matter of degree, or not otherwise localized in time. Even in biological problems, some events (for example, heart attack or other organ failure) may have the same ambiguity. The theory outlined below assumes well-defined events at specific times; other cases may be better treated by models which explicitly account for ambiguous events. More generally, survival analysis involves the modeling of time to events data; in this context, death or failure is considered an "event" in the survival analysis literature- traditionally only a single event occurs for each subject, after which the organism or mechanism is dead or broken. Recurring event or repeated event models relax that assumption. The study of recurring events is relevant in systems reliability, and in many areas of social sciences and medical research. In survival analysis, we follow patients over time, until the occurrence of particular event such as death, relapse, recurrence, or some other event that represent a dichotomy. Of special interest to the practitioners of survival analysis is the construction of survival curves, which are based on the time interval between a procedure and an event.

Information from survival analysis is used frequently to assess the efficacy of clinical trials. Researchers follow patients during the trial in order to track events.

2. Problem statement

Survival models can be viewed as consisting of two parts: underling baseline hazard function, often denoted $\lambda_0(t)$, describing how the risk of event per time unit changes over time at baseline levels of covariates; and the effect parameters, describing how the hazard varies in response to explanatory covariates x_s. According to this the research problem can be summed up in the following questions that we mention in the introduction, what is the proportions of patient with liver cancer which will survive past a certain time? Of those that survive, at what rate will they die or fail? Can multiple causes of death or failure be taken in to account? How do particular circumstances or characteristics (explanatory variables) increase or decrease the probability of survival? The answers of these question deals us to objective of the study

3. Objectives of the study

The paper aims to use the Cox model to study survival data for patient with liver cancer so as to predict Cox model of liver cancer.

4. Importance's of the study

The importance of this study is stems from provides health planners with survival information for patients with liver cancer.

5. Limitation of the study

The limitation study is the liver cancer data of military medical hospital for year 2016

6. The Study's hypotheses

H_0 : Independent variables (age, gender, housing status) not affect the time that precedes the event.

H_1 : Independent variables (age, gender, housing status) affect the time that precedes the event.

7. Study variable

The dependent variable is the time that passes until the event occurs; the explanatory variables are the gender, age and housing status.

8. Literature review

Cox described the proportional hazard model in JRSSB (1972), in what are now the most quoted statistical papers in history. He also out lined in this paper the method for estimation which he referred to as using conditional likelihood. It was pointed out to him in the literature that what he proposed was not conditional likelihood and that there may be some flaws in his logic [2]. Cox (1975) was able to recast his method of estimation through what he called "partial likelihood" and published this in Biometric. This approach seemed to be based on sound inferential principles [3].

Al-Kafrani (2017) used the Cox model as an alternative to the logistic regression model. His study focused on the Cox model because it was considered one of method used in survival analysis and has several advantages the most important is the ease of dealing with the censor data [4].

9. Methodology

The paper was based on theoretical approach that dealt with the Cox model (proportional hazards) in survival analysis data and supported the practical side that depends on the data of liver cancer from the Ministry Heath of Khartoum State (Military Medical Hospital). The paper used the SPSS for analyzing data.

10. Theoretical formulation

10.1 Model Survival Analysis

Models for analysis of data which have three main characteristics:

a- The dependent variable or response the waiting time until the occurrence of a well-defined event

b- Observation are censored, in the sense that for some units the event of interest has not occurred at the time the data are analyzed

c- There are predictors or explanatory variables whose effect on the waiting time we wish to assess or control, where the point of the event is known as the event time (survival, death, success, or failure, …etc.) its appear after a period of time where it is long or short, and the time before the event called the time of survival [5].

10.2 Survival function

We will assume for now that T is a continuous random variable with probability density function (p .d .f)$f(t)$and cumulative distribution function (c. d. f) $F(t) = P_r\{T < t\}$, giving the probability that the event has occurred by duration. It will often be convenient to work with the complement of the c.d.f, the survival function:

$$s(t) = P_r\{T \geq t\} = 1 - F(t) = \int_t^{\infty} f(Z)dZ \qquad (1)$$

This gives the probability of being a live just before duration t, or more generally, the probability that the event of interest has not occurred by duration t.

10.3 Hazard function

An alternative characterization of the distribution of T is given by the hazard function, or instantaneous rate of occurrence of the event, define as :

$$\lambda(t) = \lim_{dt\to 0} \frac{P_r\{t \leq T < t+dt / T \geq t\}}{dt} \qquad (2)$$

The numeration of this expression is the conditional probability that the event will occur in the interval $[t, t + dt]$ given that it

has not occurred before, and the denominator is the width of the interval. Dividing one by the other we obtain of interval goes down to zero we obtain an instantaneous rate of occurrence. The conditional probability in the numeration may be written as the ratio of the joint probability that T is in the interval [t, t+dt] and $T \geq t$ (which is, of course, the same as the probability that t is in the interval), to the probability of the conditional $T \geq t$. the former may be written as f (t) dt for small dt, while the latter is s (t) by definition. Dividing by dt and passing to the limit gives the useful result

$$\lambda(t) = \frac{f(t)}{s(t)} \qquad (3)$$

Which some authors give as a definition of the hazard function. In words, the rate of occurrence of the event at duration t equals the density of events at t, divided by the probability of surviving to that duration without experiencing the event. Note from the Equation (1) that –f (t) is the derivative of s(t) this suggests rewriting Equation (3)

$$\lambda(t) = \frac{-d}{dt} logs(t) \qquad (4)$$

if we now integrate from 0 to t and introduce the bounding $s(0) = 1$ (since the event is sure not to have occurred by duration 0), we can solve the above expression to obtain a formula for the probability of surviving to duration as a function of the hazard at all duration up to t:

$$s(t) = \exp\{-\int_0^t \lambda(z)dz\} \quad (5)$$

This expression should be familiar to demographers the integral in curly brackets in this equation is called cumulative hazard (or cumulative risk) and is denoted

$$A(t) = \int_0^t \lambda(z)dz \qquad (6)$$

You may think of A (t) as a sum of the risks you face going from duration 0 to t. these results show that the survival and hazard function provide alternative but equivalent characterization of the distribution of T. Given the survival function, we can always differentiate to obtain the density and then calculate the hazard using Equation (3). Given the hazard, we can always

integrate to obtain the cumulative hazard and the exponentiation to obtain the survival function using Equation (4).

An example will help fix ideas a constant risk over time, so the hazard is

$$\lambda(t) = \lambda \qquad (7)$$

for all t. the corresponding survival function is

$$s(t) = exp\{-\lambda t\} \qquad (8)$$

This distribution is called the exponential distribution with parameter λ. The density may be obtained multiplying the survivor function by the hazard to obtain

$$f(t) = \lambda exp\{-\lambda t\} \qquad (9)$$

The mean turns out to be $1/\lambda$. This distribution plays a central role in survival analysis, although it is probability too simple to be useful in application in its own right.

11. Cox regression model

The regression method knows as Cox regression (after D. R Cox who first proposed the method) these additional regression techniques are available when the dependent measures may consist of a mixture of either time- until-event data or censored time observations. We describes this technique by first introducing the hazard function, which describes the conditional probability that an event will occur at a time just larger than $t_i$conditional on having survived event-free until time t_i this conditional probability is also known as the instantaneous failure rate at time t_i and is often written as the function $h(t_i)$. The regression model requires that we assume the covariates have the effect of either increasing or decreasing the hazard for a particular individual compared to some baseline value for the function. In clinical trial if we might measure k covariates on each of the subjects where there are $I = 1, \dots, n$subjects and $h_0(t_i)$ is the baseline hazard function. We describe the regression model as:

$$h(t_i) = h_0(t_i)\exp(\beta_1 z_{i1} + \beta_2 z_{i2} + \cdots + \beta_k z_{ik} \qquad (10)$$

The regression coefficients represent the change in the hazard that results from the risk factor, z_{ik}that we have measured. Rearranging the above equation shows that the exponentiated coefficient represents the hazard ratio or the ratio of the conditional probabilities of the event. This is the basis for naming this method proportional hazards regression. Daniel [6]

$$\frac{h(t_i)}{h_0(t_i)} = \exp(\beta_1 z_{i1} + \beta_2 z_{i2} + \cdots + \beta_k z_{ik}) \qquad (11)$$

11.1Important characteristic of Cox model are that

a. It is a product of a function in t and a function in z_j
b. Z is time independent
c. The baseline hazard is an unspecified function, making it a semi- parametric model

11.2Coxes-Model Assumption

a- Non – proportional hazard

The proportional hazard s assumption is so important to Cox regression that we often include it in name (the Cox proportional hazards model). What it essentially means is that the ratio of the hazard for any two individuals is constant over time. They are proportional. It involves logarithms and it is strange concept

b- Nonlinear covariate relation ships

The Cox model assumes that each variable makes a linear contribution to the model, but sometimes the relationship may be more complex

c- Lack of independence

Lack of the independence is not something that we have to wait to diagnose unit our data is collection there are several ways to an account for of independence, but this is one problem we don't want to ignore. An invalid model will ruin all our confidence intervals and p-value.

11.3The popularity of the Cox model

a. The Cox model is robust
b. The estimated hazards are always non-negative

c. TheB_i's can be estimated and the hazard ratio calculated.
d. The h (t, z) and s (t, z) can be estimated
e. The Cox model is preferred over the logistic model which ignores survival time and censoring information.

11.4 Estimating Cox model parameter

Cox proposed the method of partial likelihood to estimate parameter of the model in the absence of knowledge of the nature of the baseline function $h_0(t)$ as this method corresponds to least square method which use in linear regression and also corresponds to the method of maximum likelihood which use in logistic regression. To estimate B_i let $t_1 < t_2 < \cdots < t_m$ refer to the time known and direction of observation and R_i refer to risk group at t_i . We assumed that only one observe has known at time $t_i$called j_i(no two events can occur at the same time)

$$L = \prod_{i=1}^{m} \frac{e^{\grave{z}J(\grave{\imath})\beta}}{\sum_{j\in R_i} e^{zJ(\grave{\imath})\beta}} \qquad (12)$$

$$\log L = \sum_{j=1}^{m} [\grave{z}_{j(i)}\beta - \log \sum_{j\in R_i}^{m} e^{z\beta_{j(i)\beta}}$$

By taking the first derivative toβ we find

$$\frac{d\log L_i}{d\beta_r} = z_{j(i)r} - \frac{\sum_{j\in R_i} e^{z_j\beta} zjs}{\sum_{j\in R_i} e^{\grave{z}j\beta}} \qquad (13)$$

The previous equation can be written as follow:

$$U_r(\beta) = \frac{d\log L_i}{d\beta_r} = z_{j(i)r} - A_{ir}(\beta) \quad \text{r = 1, 2…, p}$$

Where A_{ir} it is the weighted average $z_r$for each risk group R_i and taking the second derivative

$$I_{rs}(\beta) = \frac{\partial^2 \log L1}{\partial\beta_r \partial\beta 2} = \frac{\sum_{j\in R_i} e^{\grave{z}\beta} zjszjs(e^{\grave{z}j\beta}) - (\sum_{j\in R_i} e^{\grave{z}j\beta} zjr)(\sum_{j\in R_i} e^{\grave{z}j\beta} zjs)}{(\sum_{j\in R_i} e^{z\grave{j}\beta})^2}$$

r, s = 1, 2…p (14)

11.5Parameters significance test of the Cox model

The Wald test is used to examine the effect of independent variable in the Cox model so as the model contains a binary variable and some independent variables each one accompanied by one parameter. The Wald test takes the null hypothesis as zero

$$H_0: \beta_i = 0$$

Where the value of the test is a square of the t-test the formula is as follow

$$W_j = \left(\frac{\beta_i}{s.E\beta_i}\right)^2 \quad (15)$$

The test statistics is distributed according to chi square with one degrees of freedom in process of constructing of the model. The test leads us to know the un desirable variables which should be excluded from the model. The confidence intervals of the estimated Cox model parameter are given in the following formula

$$\hat{\beta} \pm Z\alpha_{/2}\hat{s}_E(\hat{\beta}) \quad (16)$$

11.6Model significance testing

The test of the model means knowing the fit model using the maximum likelihood ratio, this test is carried out by estimating two data models and then comparing them during the logarithm in the case of the possibility of the two models, since the model that lowest value is the best. We must know that this difference is significance using the follow formula

$$LR = -2\log\left(\frac{L_m}{L_o}\right) = 2\log L_0 - 2\log L_m \quad (17)$$

12. Application

Is this section we apply to survival data of the liver cancer using Cox model with sample data obtained from the ministry of health in Khartoum state (military medical hospital) in year 2016. The sample data in table (1) bellow

Table (1) sample patient of liver cancer

No	Sex	Time	Age	Status	Housing case
1	1	43	30	0	1
2	1	10	55	0	2
3	1	12	31	0	1
4	2	19	72	0	1
5	2	8	65	0	1
6	1	3	63	0	1

7	1	20	64	0	1
8	1	27	65	0	1
9	1	10	70	0	2
10	1	6	47	0	1
11	1	9	30	0	1
12	1	23	75	0	2
13	1	26	41	0	1
14	1	30	27	0	1
15	2	20	80	0	1
16	1	10	70	0	1
17	2	4	85	1	1
18	1	5	80	1	1
19	1	21	65	1	2
20	1	4	32	1	1
21	1	6	73	1	1
22	2	11	75	1	1
23	2	13	26	0	1
24	2	5	36	0	2
25	1	13	85	0	1
26	2	12	60	1	1
27	1	10	40	1	1
28	1	5	68	1	1
29	1	5	68	1	1
30	1	1	69	1	1
31	1	37	68	1	1
32	1	4	76	1	1
33	1	1	75	1	1
34	1	11	82	1	2
35	1	3	3	0	2
36	1	11	4	0	1
37	2	10	8	0	2
38	1	2	1	0	1
39	1	7	5	0	2
40	1	10	5	0	1
41	1	26	5	0	1
42	2	6	6	0	1
43	2	7	12	0	1
44	2	15	50	0	1

There for 1=fame, 2= female, 0=censor, 1=civil, 2= urban, 0= censor, 1=dead
Sources data: Military Medical hospital

The patient sample distributed according to status cases as the table (2)

Table2. Case processing

	N	Percent
Even	15	34.1%
Censored	29	65.9%
Total	44	100.0%

The death ratio in sample is 34.1% there for the ratio of patients whom live 65.9% .For estimating we use the partial likelihood in the table (3) as follow:

Table3. Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
Sex	-.200-	.620	.103	1	.748	.819
A	.044	.017	6.772	1	.009	1.045
L	-.764-	.796	.922	1	.337	.466

From the above table we formulate the model as follow:

$$\frac{h(t_i)}{h_0(t_i)} = \exp .819z_1 + 1.045z_2 + \cdots + .466z_3$$

12.1Model significance testing

To test the model Significance, look at the table (4) bellow

Table4.Tests of Model Coefficients

	Overall (score)			Change From Previous Step			Change From Previous Block		
-2 Log Likelihood	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
84.314	8.441	3	.038	10.542	3	.014	10.542	3	.014

We note that the probability value of maximum likelihood testis less than 5% this indicates the estimated model is significance.

12.2 Estimating and interpreting parameters of the model

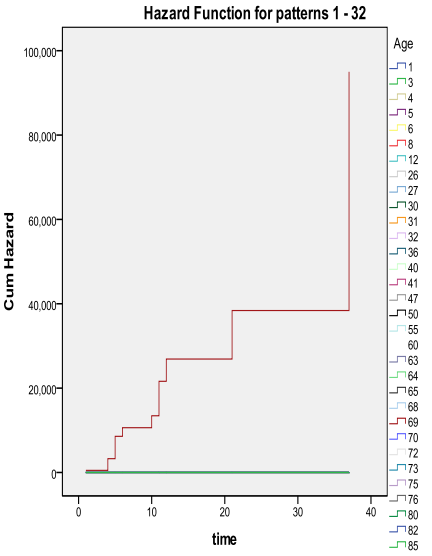
Table (3) shows the estimated parameters, their standard deviation, Wald statistic, corresponding probabilistic value, risk rate, and can explained as follows:

1- The sex variable coefficient is negative value (-0.200), the reference category adopted is the male category, and the Wald probability (0.748) is greater than 5% that means the sex variable is the not significant effect sense that risk of death in not different in female.

2- The age variable coefficient is positive and it is value is ($\beta_2 = 0.044$) this indicates that increases age by one year the risk will be increased by (1.045) times, and the probability value of Wald (0.009) is less than 5% this indicates that the age variable has a significant effect on the time of the event.

3- The housing case variable coefficient value ($\beta_3 = 0.764$) and the Wald probability value (0.337) is more than 5% this indicates that the housing case variable does not have a significant effect

12.3Estimation of the risks function of the age



We note that the risk death rate increases with age

14. Conclusion

I n this paper a Cox regression model of liver cancer data was constructed, the model includes descriptive variables such as gender, housing case variable and quantitative variable, such as age variable, and the risk rate was studied, the results were as follows:

1- Significant of the variable age and the non- significant gender variable
2- The number of deceased is less than the number of those who not die.
3- The death rate increased by age

15. Recommendations

Based on the result, the paper recommends the following

1- We recommend additional variables such as a treatment variable and social status.etc
2- The possibility of using the Cox for multiple relative risk in calculating the risk function at any given time

REFERENCES

[1] Klein Baum, D, and Klein, M., (2005), "survival Analysis: a self-learning Text", P 4, USA: Springer.

[2] Cox, D.R (1972), "Regression Models and life-tables (with discussion). J. R. Statist. 34, 187-220

[3] Cox, D. R, (1975), "partial Likelihood- Biometric 62, 269- 76

[4] Al-Kafrani (2017), "The appropriate Regression Model in Survival Analysis: in case of binary variable when taken time in consideration", Dar Almandumah, P. 75-88

[5] Kalbfleisch& Prentice (2002), "Analysis of Failure Time Data", 267-278

[6] Daniel, Wayne. W, (2005), "Bio statistics A foundation for Analysis in the Health Sciences" eighth edition, John Wiley and sons, United States of America.