Journal of Statistics and Management Systems

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tsms20

Survival analysis by using with Cox regression model for cancer patients in Basra (Model of the educational hospital)

Sahera Hussein Zain Al–Thalabi

To cite this article: Sahera Hussein Zain Al–Thalabi (2021): Survival analysis by using with Cox regression model for cancer patients in Basra (Model of the educational hospital), Journal of Statistics and Management Systems, DOI: 10.1080/09720510.2020.1859805

To link to this article: https://doi.org/10.1080/09720510.2020.1859805

Published online: 30 Mar 2021.

Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
DOI : 10.1080/09720510.2020.1859805

Taylor & Francis
Taylor & Francis Group

Survival analysis by using with Cox regression model for cancer patients in Basra (Model of the educational hospital)

Sahera Hussein Zain Al–Thalabi
Department of Statistics
Faculty of Administration & Economics
University of Basra
Basra
Iraq

Abstract

This study aims to know the factors affecting the death probability of cancer patients in Al-Sadr General Hospital and to determine the most important variables affecting the level of risk. Cox regression models were applied to find out the risk factors that significantly affect the risk function during a period of time which includes the study of time since Diagnosis of the disease until the emergence of the event (death) or observation. The study reached of we see the marital status is the only variable affecting the survival time, it is significant at a level of significance less than the level of significance preset according to the Wald test, so the equation of final regression for Cox's is It depends only on the social status variable. As such, the Moral and Social Status affect the survival time for patients with cancer, the risk condition decreases if the patient is married and his condition is for the better.

Subject Classification: *62G08; 62J02.*

Keywords: *Cox proportional hazards model, Likelihood ratio test, Score test, Wald test.*

1. Introduction

The impact study of independent variables on the dependent variable taking into account the time, The motivation and the basis for the beginning of studies and research related to the survival time of a person when a disease, and given the importance of the subject of survival time and its impact on multiple factors, has emerged the urgent need for developing statistical methods to increase the accuracy and comprehensive knowledge

E-mail: saherazain@gmail.com

of the factors that affect the survival of the injured alive or dead during the study period. One of these methods is appropriate for the status of the dependent variable that's binary response which is the most commonly used Cox regression model. This model aims to identify the risk factors that significantly affect the risk function over a period of time, which includes the study of time from the diagnosis of the disease until the emergence of the event (death) or surveillance. The prevalence of cancer in the city of Basra is remarkable and calls for the study of this disease and the application of a model appropriate to the situation, as well as the lack of studies related to Basra in particular or in Iraq in general, whether those studies are medical or statistical, and the hypothesis was based on that there is a statistical significance between the causative factors and the incidence of cancer, which was studied as a case. The aim of this study is to identify the factors affecting the (death probability) of cancer patients in the general chest hospital and to identify the most important variables affecting the level of risk. Find on mean survival time for people with cancer and come up with a model that helps to know identify cases with an increased risk of death.

Used the Cox modelto study the impact of socio-economic factors on age at first marriage in women in western Uganda [1]. The study found that the level of education, occupation, age, and area of residence are factors that significantly affect the age at first marriage. A team of researchers [3] examined a study on breast cancer in a sample of women as the sample size was (41440) a woman who was followed for two years using the semi-medial Cox model in addition to the Cox model. The study concluded factors in the event of recovery from the disease. Examined the Cox proportional risk model procedure using the concept of survival and death over time, and there is a clear relationship to how the risk of death affects over time, and the Cox model is solved by coordinated landing to save arithmetic time in solving the probability of partial record conditions for eliminating features is explored. [5]

The current study dealt with two aspects, the first related to the theoretical side of the statistical method used, and the second related to the applied side to analyze data using descriptive statistics and the Cox regression model.

2. Cox Proportional Hazards Model

This method is taking into account time as the dependent variable consists of two parts (binary descriptive variable plus time-variable) and

time is a key factor in the analysis of the phenomenon under study**.** The main advantage of this method is to study the relationship between the time leading up to the event with one or more independent variables regardless of the nature of these variables in terms of being quantitative, descriptive or mixed [9]. The formula for the Cox regression model at time t is as follows: [8] whereas:

$$\lambda(t,z) = \lambda_0(t)\exp(\underline{\beta}\underline{Z}') = \lambda_0(t)\exp\left(\sum_{i=1}^{k}\beta_i Z_i\right) \quad (1)$$

$\lambda_0(t)$: The initial function is when values (z = 0) are non-negative and unknown.

$\underline{\beta}$: Represents a vertical vector with a rank (k * 1) of regression parameters and is unknown.

($\underline{z}$) : Is a row vector with a rank (1 * k) of explanatory variables.

$\exp(\underline{\beta}\underline{Z}')$: Relative risk function that is not time-dependent, viz the effect of the explanatory variables increases or decreases the risk is constant does not change depending on the change of time point.

The ratio between any two risk ratios is constant and is not time-dependent and according to the following formula:

$$\frac{\lambda(t \setminus z_1)}{\lambda(t \setminus z_2)} = \frac{\lambda_0(t)\exp(\beta Z_1')}{\lambda_0(t)\exp(\beta Z_2')} = \frac{\exp(\beta Z_1')}{\exp(\beta Z_1')} = e^{\beta(Z_1'-Z_2')} \quad (2)$$

The formula (1) is expressed in accordance so formula (3) as follows:

$$S(t,z) = S_0(t)\exp\left[\sum_{i=1}^{k}\beta_i Z_i\right] \quad (3)$$

3. Hypotheses of the Cox Regression Model: [2]

1. The same risk function is for items that have the same values as explanatory variables.
2. When the values of a specific explanatory variable differ for two items, there will be two parallel risk functions for these two terms, known as the proportional hypothesis.

4. Estimate Model Parameters:

Due to the lack of knowledge of the t distribution, the parameters of the Cox model are estimated by finding the greatest possible function, depending on the order of the event (death) instead of the event distribution, and this is done using:

1. Method of Marginal likelihood
2. Method of Partial likelihood

The partial maximum feasibility method will be adopted in our current study because it is a common method used in the Cox model, as this method does not depend on the form of the function $\lambda_0(t)$ or any assumptions about it due to the difficulty of estimating the function, but rather depends on the arrangement of times, not on the values of actual times. For unique failure times, is increasingly required $t_1 < \bullet\ \bullet\ \bullet < t_m$ the partial probability formula for the Cox model is as follows: [10]

Rt : (the risk set) is the set of indices of samples for death or times observation occurring after (t)

h : just before time (t).

$LL(\beta) = \log\ L(\beta)$: the log partial likelihood

$$U_k(\beta) = \frac{\partial}{\partial \beta_k} LL(\beta) = \sum_{t_i} \left[z_{jik} - \frac{\sum_{h \in R(t_i)} Z_{hk} \exp(\beta Z'_h)}{\sum_{h \in R(t_i)} \exp(\beta\ Z'_h)} \right] = 0 \qquad (7)$$

$$V = -\left[\frac{\partial^2}{\partial \beta_i \beta_j} \log L(\beta) \right] = -\sum_{t_i} \left[\frac{\sum_{h \in R(t_i)} Z_{hi} Z_{hj} \exp(\beta Z'_h)}{\sum_{h \in R(t_i)} \exp(\beta\ Z'_h)} - \frac{\sum_{h \in R(t_i)} Z_{hi} \exp(\beta Z'_h)}{\sum_{h \in R(t_i)} \exp(\beta\ Z'_h)} * \frac{\sum_{h \in R(t_i)} Z_{hj} \exp(\beta Z'_h)}{\sum_{h \in R(t_i)} \exp(\beta\ Z'_h)} \right] \qquad (6)$$

Where $k = 1, 2, 3, ..., p$. in large samples.

In general, the regression equation parameters are $\beta_1, \beta_2, \cdots, \beta_p$ to be estimated And the method of Newton - Raphson is according to the following:

$$\hat{\beta}_{k+1} = \hat{\beta}_k - (V_k)^{-1} U_k(\beta) \tag{8}$$

5. Statistical Tests and Confidence Intervals:

We may want to test whether the null hypothesis $\beta = \beta_0$ is true, where usually $\beta = 0$. Different test statistics could be used. There are three typesare the most widely adopted in Cox regression:

5.1 The Wald Test Statistic (W): [7]

Test for the individual hazard ratio is based on theWald test which is testing whether the individual hazard coefficient is zero or not with:

$$H_0 : \beta_j = 0$$

Wald test is:

$$W = \left(\frac{\beta_j}{s.e(\beta_j)} \right)^2 \tag{9}$$

the Wald test Subject to to chi- square distribution with (1) degree of freedom

5.2 *Likelihood Ratio Test Statistic [4]*

The test is named as a ratio rather than a difference since the difference between two log-likelihoods is equal to the log of the ratio of the two likelihoods.

The likelihood ratio is defined as

$$LR = -2(LL_{\text{subset}} - LL_{\text{full}}) \tag{10}$$

Where:

LL_{full} : The partial weighting algorithm of a model that contains all of the variables.

LL_{subset} : The partial weighting algorithm of a model that does not contain all the variables.

The LR test Subject tochi-square distribution with (p) degrees of freedom, it can be used to obtain probabilistic values (P-value) to test for significant parameters, it is also used to test the estimated model.

5.3 *Score Test Statistic (SC): [6]*

It is a test equivalent to a logarithm rank test, the Score test is defined as:

$$SC = U'V^{-1}U$$

Where : $U = \frac{\partial}{\partial\beta}\log L(\beta)$ denotes the vector of score functions

$$V = -\left[\frac{\partial^2}{\partial\beta_i\beta_i}\log L(\beta)\right]$$

The degree test approaches the distribution of a square with a degrees (p) of freedom.

The confidence interval $(1-\alpha)\%$ of the COX regression coefficient can be obtained at the relative risk:

$\hat{\beta}_j \pm \left|Z_{\alpha/2}\right| s.e(\hat{\beta}_j)$ is the two-sided confidence interval. [10]

6. Statistical Analysis

The study adopted a sample from the records of Al-Sadr General Hospital in Basra for people with cancer, and the sample reached (132) patients, and this sample was dependence on some factors that directly or indirectly affect the patient's death, such as (the patient's age, the patient's sex, marital status, area housing well as the location of the tumor). The SPSS program was fed with data for the study sample to analyze the data statistically using descriptive statistics methods along with the Cox regression model [11].

6.1 *Description of the Study Variables*

The variables of the study were determined based on the factors that have the main role in determining the seriousness of the disease and the available and recorded in the above-mentioned hospital records.

a. ST Time : (Survival Time) It represents the dates of diagnosis of the disease until the date of death or the date of the last follow-up of the patient during the study period.
b. Status variable : takes the value (1) when the patient dies and takes the value (0) if the patient is alive or lost follow-up.

(Z_1) : Gender variable: if Male = 1and if Female = 0.

(Z_2) : Age variable: The patient's age was divided into three categories according to the following: First category: (15-25) years = 1, second category: (26-40) years = 2 , third category: (41 and more) = 3 .

(Z_3) : Social Status Variable: It takes the value (1) if the patient is married and takes the value (0) if the patient is otherwise.

(Z_4) : The location of the residence: It was classified according to the county districts and according to the following:Basra Center=1, Zubair district=2, spend Al-Midaina =3, Shatt al-Arab District=4, Qurn district=5, Fao district=6, spend Abu Al-Khaseeb = 7

(Z_5) : The site of the tumor: It was classified according to the presence of the tumor in the patient's body organ and the study sample, according to the following:Breast = 1, Respiratory system = 2, Digestive system = 3, Lymphatic system = 4, The nervous system (brain) = 5

6.2 *Estimation of Parameters Cox Regression:*

The Cox model parameters were estimated as five variables (patient sex, age, marital status, area of residence, and tumor location) have an effect on the riskfunction of cancer patients, as in equation (1) as follows:

We note in Table (1) that parameter (β_2) is positive, indicating an increase in one unit of the explanatory variable (Z_2) will lead to an

Table 1

Results Cox- Regression Model Estimation according to the Enter

Variables	β	SE	Wald	df	Sig.	The limits of confidence (95.0%)	
						Lower	Upper
Z_1	-.232	.297	.612	1	.434	.443	1.419
Z_2	.373	.213	3.057	1	.080	.956	2.204
Z_3	-.912	.316	8.348	1	.004	.216	.746
Z_4	-.103	.082	1.578	1	.209	.769	1.059
Z_5	-.096	.101	.897	1	.344	.745	1.108

Source : From the prepare by the researcher based on the program SPSS

increase in the risk function and that the case under study is heading for the worse, while the parameters (β_1, β_3, β_4, β_5) for the variables (Z_1, Z_3, Z_4, and Z_5) respectively, so they are negative and this indicates that when increasing one unit of any of the variables will lead to a decrease in the risk function, i.e. the case under study is heading towards a better direction. The table indicates the significance of the variable (Z_3), through the value of the Wald test as the probability value corresponding to the test value is (0.004) and is less than the value of the predefined significance level (0.05), which means that the Moral and Social Status affect the survival time for patients with cancer, and indicate the negative value of the variable (β_3 = -0.912) that the patient when it is married to decrease the risk function and heading towards a better direction by exp (-0.912).

Used of the reverse test method to determine the best regression equation and the results were summarized in the tables as follows:

Table 2

Results Cox- Regression model estimation according to the method for Backward

		β	SE	Wald	df	Sig.	The limits of confidence (95.0%)	
							Lower	Upper
Step 1	Z1	-.232	.297	.612	1	.434	.443	1.419
	Z2	.373	.213	3.057	1	.080	.956	2.204
	Z3	-.912	.316	8.348	1	.004	.216	.746
	Z4	-.103	.082	1.578	1	.209	.769	1.059
	Z5	-.096	.101	.897	1	.344	.745	1.108
Step 2	Z2	.359	.214	2.810	1	.094	.941	2.180
	Z3	-.957	.312	9.422	1	.002	.208	.707
	Z4	-.094	.081	1.346	1	.246	.776	1.067
	Z5	-.120	.098	1.503	1	.220	.733	1.074
Step 3	Z2	.318	.209	2.321	1	.128	.913	2.070
	Z3	-.948	.310	9.359	1	.002	.211	.711
	Z5	-.123	.097	1.607	1	.205	.731	1.070
Step 4	Z2	.285	.207	1.900	1	.168	.887	1.993
	Z3	-.857	.301	8.119	1	.004	.235	.765
Step 5	Z3	-.727	.282	6.627	1	.010	.278	.841

Source: From the prepare by the researcher based on the program SPSS

Table (2) shows the steps of the inverse regression and the significance of the parameters in each step according. Note that in the first step, we see that the first variable, Z1, has the lowest value for the Wald statistic, and it is not significant at the pre-determined significance level (0.05), that is, it has no effect on the survival time, so it is removed from the model. After that we repeat this steps to remove the variable that is not significant, so we remove the variables Z4 , has the lowest statistic the second step , Z5 in the third step, and the variable Z2 in the fourth step After removing the variables one by one due to their lack of significance, we see in the fifth step that the variable Ð3 (marital status) is the only variable affecting the survival time, it is significant, we see in the fifth step that the variable Z_3 (marital status) is the only variable affecting the survival time, it is significant at a level of significance less than the level of significance preset (0.05) according to the Wald test, so the form the final regression for Cox's is as follows:

$$Y = \lambda_0(t)\exp(-0.727)Z_3$$

$$Y = Ln\frac{\lambda(t / Z_1, Z_2, Z_3, Z_4, Z_5)}{\lambda_0(t)} = (-0.727)Z_3$$

The best model for the study was determined through the Likelihood Ratio, as shown in Table (3).

We notice in Table (3) that the model in the first step that includes all the variables is less significant than the models in the steps by comparing the value of (Sig.) with the predetermined level of significance (0.05), as the table shows that the model containing the social condition factor in the

Table 3

Results of Test for Best Model with using likelihood ratio statistic

Step	-2 L L	Overall (score)		
		Chi-square	df	Sig.
1[a]	417.564	12.041	5	.034
2[b]	418.185	11.383	4	.023
3[c]	419.635	10.431	3	.015
4[d]	421.259	8.695	2	.013
5[e]	423.210	6.897	1	.009

Source: From the prepare by the researcher based on the program SPSS

step the fifth is the most significant model from the rest of the models in the previous steps.

Conclusions

The study reached if we see the variable Z_3 (marital status) is the only variable affecting the survival time, it is significant at a level of significance less than the level of significance preset (0.05) according to the Wald test, so the equation of final regression for Cox's is depends only on the social status variable.

The moral and social status affect the survival time for patients with cancer, the risk condition decreases if the patient is married and his condition is for the better.

References

[1] Agaba, P. , Atuhaire, L. and Rutaremwa, G., 2010. Determinants of age at first marriage among women in western Uganda, *Department of PopulationStudies*, SSAE, Makerere University, Uganda.

[2] Bakir, E., Ramadan, I., 2015. Using Statistical Methods to Study the Determinants of Marital Status for Women in Palestine "Comparative Applied Study, Master Thesis in Statistics, *Department of Applied Statistics, Faculty of Economics and Administrative Sciences,* Al-Azhar University, Gaza.

[3] Chlebowski , R. and others, 2013. Estrogen Plus Progestin and Breast Cancer Incidence and Mortality in the Women's Health Initiative Observational Study, *Journal of National Cancer Institute - Oxford University*, Vol. 105, No. 8, pp. (526-535)

[4] Hosmer, D., Lemeshow, S. and May, R., 2007. Applied Survival Analysis: Regression Modeling of Time to Event Data, 2nd edition Wiley, New York, USA.

[5] Jessica K., 2017. Solving the Cox Proportional Hazards Model and Its Applications, Master of Science, the *Department of Electrical Engineering and Computer Sciences*, University of California, Berkeley.

[6] Kalbfleisch, D. and Prentice, L., 2002. The Statistical Analysis of Failure Time Data, 2nd ed. Wiley, New York.

[7] Klein, J. and Moeschberger, M., 2003. Survival Analysis Techniques for Censored and Truncated Data, 2nd ed., Springer.

[8] Mohammed, M., 2014. Survival Analysis by Using Cox Regression Model with Application, *International Journal of Scientific & Technology Research,* Volume 3, No.11, p:316

[9] Talabani, S., 2012. Comparative study between logistic regression models and Cox regression model to study the most important economic and demographic factors affecting the knowledge and attitudes of young people towards reproductive health issues, unpublished doctoral thesis, Abubakar Belgard University, Tlemcen, Algeria.

[10] Zhang, H., 2015. Checking proportionality for Cox's regression model, Master Modelleringog data analyzed, Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Oslo.

[11] Shubham Sharma & Ahmed J. Obaid (2020) Mathematical modelling, analysis and design of fuzzy logic controller for the control of ventilation systems using MATLAB fuzzy logic toolbox, *Journal of Interdisciplinary Mathematics,* 23:4.

Received November, 2020

Revised January, 2021