The Scientific Journal of Cihan University – Sulaimanyia **PP: 94-108**
Volume (3), Issue (1), Jun 2019
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Predicting Gastric Cancer Survival Patients Using Cox Regression Model

Professor Dr. Monem A. Mohammed[1], Shokh Mukhtar[2]

[1,2]Statistics and Informatics Department, College of Administration and Economy, University of Sulaimani, Iraq
Monem_aziz2003@yahoo.com[1]
shoch.ahmad@freenet.de[2]

Abstract

Cox Regression model is one of the important models that can be used to analysis the survival data, and can detect relationship between the explanatory variables and their survival time. The Cox Regression model is a semi-parametric model that composed of two parts: first part is non-parametric ($\boldsymbol{\lambda_0(t)}$) and the second is parametric part ($\mathbf{exp}$ ($\underline{\grave{\boldsymbol{\beta}}}\ \underline{\boldsymbol{z}}$)). In this study Cox-Regression model is used to predicting the survival time of patients that are suffering from Gastric cancer in Sulaimani City each of (Hemoglobin (Hb), Weight and Number of Chemotherapy) variables are have effect on survival time in this study. That have been taken for the patients of age (46 to 63) years old. The data that we have used in the study is left-censored. First, after testing distribution of survival time by using goodness of test, we noticed that the distribution of survival time is unknown. So that selecting Cox Regression as the best model to analysis stomach cancer data. Next, we checked the assumption of Cox Regression graphically by using Kaplan-Meier estimator, to estimate the survival function from lifetime data of patients. Finally, we estimate the parameters using Partial likelihood method and tested the model parameters by (Wald) test which shows that only three parameters (Hemoglobin (Hb), Weight, Number of chemotherapy) that effecting survival time of Gastric Cancer patients.

Keywords: Survival analysis, Goodness of fit test, Cox- Regression model

الملخص

يعتبر نموذج (كوكس للانحدار) من النماذج المهمة في تحليل بيانات البقاء على قيد الحياة وكشف العلاقة بين المتغيرات التوضيحية ومتغير زمن البقاء. ولذلك فان نموذج انحدار كوكس هو من النماذج شبه المعلمية حيث يتكون في جزئين، الجزء الاول يمثل الجزء اللامعلمي ($\lambda_{0(t)}$) اما الجزء الثاني يمثل الجزء المعلمي ($\mathbf{exp}(\underline{\boldsymbol{\beta}}\ \underline{\boldsymbol{Z}})$)، عندما ($\underline{\boldsymbol{\beta}}$) يمثل متجهة المعلمات غير المعلومة، ($\underline{\boldsymbol{Z}}$) فهو يمثل متجهة المتغيرات التوضيحية. في هذه الدراسة تم استخدام نموذج انحدار كوكس لتقدير (زمن البقاء) للمرضى الذين يعانون من سرطان المعدة في مدينة السليمانية حيث تم التعرف على المتغيرات المهمة لمده الدراسة وهي (الهيموغلوبين، الوزن، عدد المعالجات الكيميائية) وتم دراسة المرضى للفئات العمرية (46 الى 63) سنة وتم استخدام بيانات المراقبة من جهة اليسار. بعد اختبار بيانات المرضى لمعرفة التوزيع المناسب لهذه البيانات وحسب اختبار (دقة وحسن الاختبار) حيث وجد ان البيانات ليس لها توزيع معلوم وعلى هذا الاساس تم استخدام نموذج انحدار كوكس المناسب لهذه البيانات وتحليل بيانات مرضى سرطان المعدة. وكذلك تم استخدام فرضيات نموذج كوكس بطريقة الرسم ل(Kaplan Meier) وتقدير دالة البقاء في بيانات زمن البقاء للمرضى وكذلك تم استخدام طريقة التقدير(**Partial likelihood**) واختبار معلمات النموذج حسب اختبار (**Wald**) اللتي اضهرت هنالك فقط ثلاثة متغيرات لها تاثير على زمن البقاء للمرضى وهية (الهيموغلوبين،الوزن،عدد المعالجات الكيميائية) والمؤثرة على سرطان المعدة.

DOI: http://dx.doi.org/10.25098/3.1.7

The Scientific Journal of Cihan University – Sulaimanyia **PP: 94-108**
Volume (3), Issue (1), Jun 2019
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

پوختە

مۆدێلی چەماوەی کۆکس یەکێکە لە مۆدێلە گرنگەکان بۆ شیکارکردنی ئەو زانیارییانەی کە بـۆ مانـەوە لـە ژیانـدا وەردەگیرێـت ، هـەروەها بـۆ دیاریکردنی پەیوەندی لە نێوان گۆڕاوی پاڵپشت کە کاتی مانەوەی نەخۆشە لە ژیاندا لەم تـوێژینەوەماندا لە گەڵ چەند گۆڕاوێکی ڕوون کەڕەوە ، ئەم جۆرە مۆدێلە لە دوو بەش پێك دێت بەشێکیان ناپارامیتەرییە ($\lambda_{0(t)}$) و بەشەکەی تریان پارامیتەرییە ($\exp(\underline{\beta Z})$) ، لەم تـوێژینەوەدا مۆدێلی چەماوەی کۆکسمان بەکارهێناوەی بۆ خەمڵاندنی کاتی لە ژیاندا مانەوەی ئەو نەخۆشانەی کە دەنالێنن بە دەست نەخۆشی شـێرپە نجەی گەدە پشت بەستن بە چەند فاکتەرێکی دیاریکراوی وەك (ڕێژەی خرۆکە سورەکانی خوێن ، کـێش ، ڕێـژەی خرۆکـە سـپیەکانی خـوێن ، تەمـەن ،ژمارەی ئەو جارانەی کە چارەسەریان وەرگرتووە) بۆ ئەو نەخۆشانەی کە تەمەنیان لە نێوان (46-63) ساڵدایە لـە نەخۆشـخانەی هیـوا لـە شاری سلێمانی . گۆڕاوی پاڵپشت تـاقیکردنەوەی بۆ کرا بۆ ئەوەی بزانین سەر بە چ تەوزیعێکی ئاماریـیە لە ئە نجامدا بۆمان دەردەکەوت کە سـەر بە هیچیان نییە بۆیە تـوانیمان مۆدێلی چەماوەی کۆکس بەکاربهێنین هەروەها کاپلان مایەر بەکارهێناوە بۆ ئەوەی بزانین ڕێـژەی مەترسـی لـە نەخۆشێکەوە بۆ نەخۆشێکیتر بە نەگۆڕی دەمێنێتەوە یان نا ، لە ئە نجامدا دەرکەوت بە نەگۆڕی دەمێنێتەوە لە کۆتایدا تـاقیکردنەوەی ،والدمان بەکارهێنا بۆ ئەوەی بزانین کام لەو پێنج فاکتەرەی کە وەرمانگرتووە کاریگەری دەبێت لەسەر مانەوەی نەخۆش لە ژیاندا ، بۆمان دەرکەوت کـە هەریەکە لە کێش و ڕێژەی خرۆکە سورەکانی خوێن و ژمارەی ئەو جارانەی کە چارەسـەریان وەرگرتـووە کاریگـەری ڕسـتەوخۆیان هەیـە لەسـەر مانەوەی نەخۆش لە ژیاندا.

1. Introduction:

Survival analysis is a valuable and are mostly common branch of statistics that deals with analysis of time to events, such as death in biological organisms and failure in mechanical systems. Survival analysis makes an attempt to analysis the proportion of a population which will survive past a certain time [1]. Clinical researches with long-term follow-up regularly measure time-to-event outcomes, like survival time, for which multivariable models are used to make prediction and identify covariate associations. The foremost common approach to model covariate effects on survival is the Cox proportional hazard model, which can handle truncated and censored observations Regression analysis is generally used to recognize the risk factors. Simple logistic regression analysis is limited of only allowing a view of survival probability over the entire study as a single time interval and it assumed that patients are at risk over the entire study period [2]. This is not valid for studies with long follow up or where patients have variable time at risk. For this logic, in survival analysis, Cox's regression model is widely applicable assumptions about the shape of the baseline hazard function. So the Cox model is sometimes referred to as a semi- parametric model.

1.1 Aim of the study

The aim of this study is to detect the effect of factors (Age, White blood cells (Wbc), Hemoglobin (Hb), Weight and number of chemotherapy) on survival time by using Cox Regression model.

1.2 Literature review

[11]: This study compared three different ways to perform variable selection in the Cox proportional model, stepwise regression, lasso and bootstrap. Study also represents how simulating

The Scientific Journal of Cihan University – Sulaimanyia **PP: 94-108**
Volume (3), Issue (1), Jun 2019
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

survival data were controlled which covariates that were significant for the response. Study evaluate how well each method performs in finding the correct model.

[12]: Study result make it clear that the Cox proportional hazards model allows data to be analyzed with a concept of survival and death overtime. It is a clear relationship of how the risk of death is affected by time and the features of the data. The probabilities of surviving past a certain time are used to predict loan defaults. Study shows that there is an understanding which characteristic correlate with survival for dogs and cats in animal shelters is also possible through creating survival curves.

[13]: This study survival analysis Approach for prostate Cancer is carried out on survival analysis. Study result showed that the metastatic tumor has a poor survival rate compared to the primary tumor, which has a hint that primary tumor has higher probability.

[14]: Study represents the proportional hazard functional regression model for data where the outcome is the possibly censored time to event and the exposure is a densely sampled functional process measured at baseline. Study result represents that the model is very malleable and modular. Flexibility of the Model can be extended to incorporate a number of advances in the fields of survival analysis and in functional data analysis.

[15]: This study used the parametric survival approach to analyze the survival time cancer patients. It's obvious from other researches about the survival time of patients, the Cox Proportional Hazard model, a semi- parametric method, is primarily used. The difference in this study is that this approach does not rely on the distributional assumptions. Study shows that the parametric method is more consistent with a theoretical approach compared to a semi-parametric approach. Study results shows that the survival time of cancer patients follows the Weibull Probability Distribution.

[16]: In this study hazard function for gastric cancer patients was estimated using Wavelet and Kernel methods and some related factors. The effect of some factors on cancer patient's survival time was assessed. From the study is clear that wavelet smoothing method works perfectly on estimating of hazard function in survival analysis.

2. Some important Definition [3],[4],[5],[6]

2.1 Survival function:

(T) is a continuous random variable with cumulative distribution function $F(t)$ in $[0,\infty)$.

Survival function is:

$$\mathbf{S(t) = Pr(T > \mathit{t}) = \int_{t}^{\infty} f(u)du = 1 - F(t)}$$

Properties:

- $S(t)$ is monotonically decreasing, $S(u) \leq S(t)$ for all $u > t$.

DOI: http://dx.doi.org/10.25098/3.1.7

The Scientific Journal of Cihan University – Sulaimanyia **PP: 94-108**
Volume (3), Issue (1), Jun 2019
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

- Time , $t = 0$, represents some origin, typically the beginning of a study
- $S(0)$ is often unity but can be less to represent the probability that the system fails immediately upon operation.

3. Lifetime distribution function and event density:

The lifetime distribution function, regularly denoted F, is defined as the complement of the survival function

$$\boldsymbol{F(t) = \Pr(T \leq t) = 1 - S(t)}$$

If ($\boldsymbol{F}$) is differentiable then the derivative, which is the density function of the lifetime distribution, is denoted ($\boldsymbol{f}$),

$$\boldsymbol{f(t) = F'(t) = \frac{d}{dt}F(t)}$$

The function ($\boldsymbol{f}$) is sometimes called the **event density** and it is the rate events per unit time.

4. Failure rate and cumulative of failure rate:

It is defined as the rate of probability of failure to probability of survival at the same period of time ($\boldsymbol{T \geq t}$).

$$\boldsymbol{\lambda(t) = Lim_{\,dt \to 0} \frac{Pr(t \leq T < t + dt \setminus T \geq t)}{dt} = \frac{f(t)}{S(t)} = -\frac{\acute{S}(t)}{S(t)}}$$

The hazard function required to be:

- non-negative, $\lambda(t) \geq 0$,
- its integral over $[0, \infty]$ must be infinite, the inverse mustn't be true.
- sometimes increasing or decreasing, non-monotonic, or discontinuous.
- hazard function can alternatively be characterized in terms of the cumulative hazard function, generally denoted by ($\boldsymbol{\Lambda}$):

$$\boldsymbol{\Lambda(t) = -log\ S(t)}$$

Transposing signs and exponentiation:

$$\boldsymbol{S(t) = exp\ (-\Lambda(t))}$$

or differentiating (with the chain rule)

$$\boldsymbol{\frac{d}{dt}\Lambda(t) = -\frac{S'(t)}{S(t)} = \lambda(t)}$$

The name "cumulative hazard function" is derived from the fact that

The Scientific Journal of Cihan University – Sulaimanyia **PP: 94-108**
Volume (3), Issue (1), Jun 2019
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

$$\Lambda(t) = \int_0^t \lambda(u)\ du$$

Which is the "accumulation" of the hazard over time.

From the definition of $\Lambda(t)$, is clear that it increases without bound as (t) tends to infinity ($S(t)$ tends to zero). Cumulative hazard has to diverge. This implies that $\lambda(t)$ must not decrease too quickly. For instance, the integral of $\exp(-t)$ converges to (1), so it is not the hazard function of any survival distribution.

5. Censoring and Truncation

A truncated observation is one that is unobservable due to a selection process inherent in the study design.

- Right truncation: Arises when the whole study population has already experienced the event of interest, such as death.
- Left truncation: Arises when the subjects have been at risk before entering the study

Censored observation occurs when the exact failure time is unknown, but can only be determined to lie within a certain interval.

- Right Censoring: Right censoring arises when the study ends before the event has occurred or when a subject leaves the study before an event occurs.
- Left censoring: Is when the event has already occurred before enrolment. This is very rarely encountered [2].

6. Multiple Regression model:

Is a model with more than one independent variable can be representing as follow:

$$\underline{Y} = \underline{\acute{\beta}}\, Z + \underline{e} \quad \text{.....(1)}$$

$\underline{Y}$: is the dependent variable.

($\underline{\acute{\beta}}$) : is unknown parameters vector.

(Z): Is a non-singular matrix of independent variables.

To check the normality plot residual versus the (Z) values and other residual diagnostics.

There are some Proportional Hazards Models for survival data as follows:

6.1 Exponential Regression model [3]:

This model should be use when the response variable is exponentially with (pdf):

$$f(t\ |\underline{z}) = \frac{1}{\lambda \underline{z}}\ exp\left(\frac{-t}{\lambda \underline{z}}\right), \qquad t > 0 \qquad \dots (2)$$

DOI: http://dx.doi.org/10.25098/3.1.7

The Scientific Journal of Cihan University – Sulaimanyia **PP: 94-108**
Volume (3), Issue (1), Jun 2019
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

$(\lambda \underline{z})$: is a constsnt rate function, Where:

$\lambda \underline{z} = E(t \mid \underline{z}) = exp(\acute{\beta}\ \underline{z})$ which depends on regression parameters $(\underline{\beta})$ and explanatory variables $(\underline{z})$.

Therefore, the survival functions is:

$$S(t \mid \underline{z}) = exp\{-(\frac{t}{exp(\acute{\beta}\ \underline{z})})\} \qquad \ldots\ldots (3)$$

So the likelihood function is the product of the likelihood of each datum as follows:

$$L(\underline{\beta}, t, \underline{z}) = \pi_{i=1}^{n} (\frac{1}{exp(\acute{\beta}\ z_i)})\ exp\ (\frac{-t}{exp(\acute{\beta}\ z_i)}) \qquad \ldots\ldots (4)$$

6.2 Weibull Regression Model [3]:

If survival time is distributed with weibull distribution, then weibull regression model must be used:

$$f(t \mid \underline{z}) = \frac{\alpha}{exp(\acute{\beta}\ \underline{z})} (\frac{t}{exp(\acute{\beta}\ \underline{z})})^{\alpha-1} exp\{\frac{-t}{exp(\acute{\beta}\ \underline{z})}\}^{\alpha}\ , t > 0\ , \alpha > 0 \qquad \ldots\ldots.(5)$$

The failure rate of weibull regression model can be computed according below equation:

$$\lambda(t \mid \underline{z}) = \frac{\alpha}{exp(\acute{\beta}\ \underline{z})} (\frac{t}{exp(\acute{\beta}\ \underline{z})})^{\alpha-1} \qquad \ldots\ldots (6)$$

Also, the survival function can be computed according below equation:

$$S(t \mid \underline{z}) = exp\{\frac{-t}{exp(\acute{\beta}\ \underline{z})}\}^{\alpha} \qquad \ldots (7)$$

Therefore, the likelihood function can take the following:

$$L(\underline{\beta}, t, \underline{z}) = \pi_{i=1}^{n}\{\frac{\alpha t^{\alpha-1}}{[exp(\acute{\beta}\ \underline{z})]^{\alpha}} exp(\frac{t}{exp(\acute{\beta}\ \underline{z})})^{\alpha}\}\{exp\ (\frac{t}{exp(\acute{\beta}\ \underline{z})})^{\alpha}\} \qquad \ldots (8)$$

7. Cox-Regression model [1],[3],[4]:

The Cox proportional-hazards model is mostly commonly models that used in medical studies to detect the association between the survival times of patients with explanatory variables.

$$\lambda(t; z) = \lambda_0(t)\ exp(\acute{\beta}\underline{Z}) \qquad \ldots\ldots\ldots\ldots\ldots\ldots (9)$$

$\lambda(t, \underline{z})$: is a hazard rate function at time **(t)** for an individual with covariates $(\underline{Z})$.

$\lambda_0(t)$: is an unspecified base- line hazard function for continuous **(t)**.

$\acute{\beta}$: is the slope coefficients.

The Scientific Journal of Cihan University – Sulaimanyia **PP: 94-108**
Volume (3), Issue (1), Jun 2019
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

The density function **S (*t*)** is:

$$S(t;\underline{Z}) = (t;\underline{z})S(t;\underline{Z})$$

$$= exp\ (- \int_0^t \lambda_0(u)exp(\acute{\beta}\ \underline{Z})du) \qquad \text{.....(10)}$$

Assumptions

1. The proportional hazard should be fixed from a patient to another in the study.
2. The natural log of the hazard function must have a linear relationship with the explanatory variables.
3. The explanatory variable should not be depending on time.
4. The response variable must not distribute any statistical distribution.
5. The hazard rate should increase linearly with time.

8. Partial likelihood method:

Cox (1975) has developed partial log-likelihood method, which is a nonparametric method to include the covariate estimates of Cox-regression model, hazard ratios can be estimated by using maximum likelihood techniques. The partial likelihood is valid first if there are no ties in the data set that mean if two subjects have not the same event time. Otherwise, the true partial log-likelihood function involves permutations and can be time-consuming to compute.

Then, to study (Cox) model which have the following hazard function [4]:

$$\lambda(t \setminus \underline{z}) = \lambda_0(t)\ exp\ (\acute{\beta}\ \underline{z}) \qquad with \quad T_1 < T_2 < \ldots < T_n \qquad \text{.....(11)}$$

Survival models can be usefully viewed as ordinary regression models in which the response variable is time. However, computing the likelihood function (needed for fitting parameters or making other kinds of inferences) is complicated by the censoring. So, the likelihood function can take the following:

$$L(\underline{\beta}, \lambda_0(t), t, \underline{z}) = \pi_{i=1}^{n}\{\lambda_0(t_i)\ exp\ (\acute{\beta}\ \underline{z})\}^{\delta_i}\ exp\ \{-\int_0^{t_i} \lambda_0(u)exp\ (\acute{\beta}\ \underline{z})du\}$$

$$= \pi_{i=1}^{n} \frac{exp(\acute{\beta}Z_i)}{\{\sum_{L\in R(t_i)} exp(\acute{\beta}Z_L)\}} \sum_{L\in R(t_i)} \lambda_0(t)exp\left(\acute{\beta}Z_L\right) \pi_{i=1}^{n} S_0(t_i)\ exp\ (\acute{\beta}\ \underline{z}) \qquad \text{..... (12)}$$

Where

$$S_0(t) = \exp\ (- \int_0^t \lambda_0(u)du\ )$$

The previous likelihood equations are special cases of (3). Equation (3) can be approximated by:

DOI: http://dx.doi.org/10.25098/3.1.7

The Scientific Journal of Cihan University – Sulaimanyia **PP: 94-108**
Volume (3), Issue (1), Jun 2019
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

$$L(\underline{\beta}, t, \underline{z}) = \pi_{i=1}^{n} \frac{\exp(\underline{\acute{\beta}}\, \underline{Z}_i)}{\{\sum_{L\in R(t_i)} \exp(\underline{\acute{\beta}} \underline{Z}_L)\}}$$

The maximum likelihood estimate of ($\underline{\beta}$) is ($\underline{\hat{\beta}}$) and can be obtained as a solution to the system of the following equations:

$$\frac{\partial \log pL(\underline{\beta}, t, \underline{z})}{\partial \beta_i} = \sum_{i=1}^{k} \{ \underline{Z}_i - \frac{\sum_{L\in R(t_i)} \exp(\underline{\acute{\beta}}\, \underline{Z}_L) \underline{Z}_{Li}}{\sum_{L\in R(t_i)} \exp(\underline{\acute{\beta}}\, \underline{Z}_L)} \}$$

and similarly, one can get:

$$\frac{\partial^2 \log pL(\underline{\beta}, t, \underline{z})}{\partial\, \underline{\beta}_i\, \partial\, \underline{\beta}_j} = \sum_{i=1}^{k} \{ \frac{\sum_{L\in R(t_i)} \exp(\underline{\acute{\beta}}\, \underline{Z}_L) \underline{Z}_{Li} \underline{Z}_{Lj}}{\sum_{L\in R(t_i)} \exp(\underline{\acute{\beta}}\, \underline{Z}_L)} - \frac{\sum_{L\in R(t_i)} \exp(\underline{\acute{\beta}}\, \underline{Z}_L) \underline{Z}_{Li}}{\sum_{L\in R(t_i)} \exp(\underline{\acute{\beta}}\, \underline{Z}_L)} * \frac{\sum_{L\in R(t_i)} \exp(\underline{\acute{\beta}}\, \underline{Z}_L) \underline{Z}_{Lj}}{\sum_{L\in R(t_i)} \exp(\underline{\acute{\beta}}\, \underline{Z}_L)} \} \quad \text{.... (13)}$$

Where (j = 1, 2, 3 … S)

9. Testing data for Goodness [5]:

9.1 Wald Test:

A common way to test for the individual hazard ratio is based on Wald test which is testing whether the individual hazard coefficient is zero or not with $H_0: \beta_j = 0$ Vs. $H_1: \beta_j \neq 0$

The Wald test (W_j) = { $\frac{\beta_j}{SE\,(\beta_j)}$ **}^{2}** .... **(14)**

9.2 Likelihood Ratio Test

Likelihood ratio tests is common and widely used, especially when comparing nested models that differ with respect to multiple parameters, proportional model fitted two models L_0 and L_1

Where:

L_0 : is the model without including any explanatory variables.

L_1 : the model with **k** including any explanatory variables.

$$LR = 2\{L_1 - L_0\} \quad \text{..... (15)}$$

Where: LR ~ X^2_m , with *m* degrees of freedom.

The Scientific Journal of Cihan University – Sulaimanyia **PP: 94-108**
Volume (3), Issue (1), Jun 2019
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

9.3 Cox & Snell R^2 Test

Cox and Snell's R^2 is based on the log likelihood for the model compared to the log likelihood for a baseline model. Even so, with categorical outcomes, it has a theoretical maximum value of less than 1, even for a "perfect" model. The Cox and Snell index is represented as [17]:

$$R^2_{CS} = 1 - \left(\frac{L(Null)}{L(Full)}\right)^{\frac{2}{n}} \qquad \text{.....(16)}$$

where:

L(Null) and **L(Full)** are the likelihood functions for the constant-only model and the model with the predictors, respectively.

n is the sample size

10. Experimental part

10.1 Introduction [7],[10]:

According to World Health Organization (WHO) Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018. Gastric cancer is the fourth most frequently occurring cancer in men and the seventh most commonly occurring cancer in women. There were over one million new cases in 2018.An estimated 783,000 deaths (equivalent to 1 in every 12 deaths Worldwide), making it the fifth most frequently diagnosed cancer and the third leading cause of cancer death. Gastric malignant growth is more common in less developing countries than in more developed countries, with about 70 per cent of cases occurring in less developed countries. The prognosis of gastric cancer is generally poor and also has a high mortality rate worldwide. Because of the low survival rate of gastic cancer patients, it is very important to discover the factors that influence survival in gastric cancer patients. A set of screening program and advances in therapy for gastric cancer patients have been shown to have contributed to the decrease in the mortality rate in such a patient.

10.2 Description of the data

This study contained 10 variables for 53 patients of ages (46 to 63) years old that suffering from Gastric cancer, the data is taken from Hiwa hospital at Slemani city:

Z_1 = Time: Spent time of patient at hospital in days.
Z_2 = Status: The status of each patient (Death or Censored).
Z_3 = Age: The Age of Patients on registration day.
Z_4 = White blood cells (Wbc): White blood cells are part of the body's immune system. They help the body fight infection and other diseases It may be used to look for conditions such as infection, inflammation, allergies, and leukemia. Measured in cell per liter (10^9/L).
Z_5 = Hemoglobin (Hb): Hemoglobin is a protein in red blood cells that carries oxygen.

DOI: http://dx.doi.org/10.25098/3.1.7

The Scientific Journal of Cihan University – Sulaimanyia **PP: 94-108**
Volume (3), Issue (1), Jun 2019
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

from lungs to the rest of body. Measured in grams per deciliter (g/dl).

Z_6 = Weight Weight of patients measured in kilogram (Kg).

Z_7 = Number of chemotherapy: Number of chemotherapy that received each patient.

Z_8 = Type of treatment: Treatments that the patients received.

Z_9 = Degree of cancer dieses: Stand for staging the cancer Type (1,2,3,4)

Z_{10} = Occupation of patients

10.3 Description and model estimation

The description and estimating the model can be shown in these below steps:

First step: Presenting the table frequency of status, type of treatment, degree of cancer dieses and occupation variables as represented below:

Table (1): Represents frequency table of status variable

Status	Frequency	Percent %
Event-Death	18	33.96
Censored	35	66.04
Total	**53**	**100**

From the above table it is clear that 33.96% of the patients under study are died and 66.04% are censored.

Table (2): Shows frequency table of type of treatment variable

Treatment	Frequency	Percent %
Drugs	2	3.77
Chemotherapy	38	71.70
Surgery	2	3.77
Drugs & Chemotherapy	1	1.89
Chemotherapy & Surgery	10	18.87
Total	**53**	**100**

Sum to up table 3.77% of patients are received drugs as treatment, 71.70% received chemical, 3.77% received surgery, 1.89% received drugs and surgery finally 18.87% received chemical and surgery, most of the patients took chemical treatment.

The Scientific Journal of Cihan University – Sulaimanyia PP: 94-108
Volume (3), Issue (1), Jun 2019
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Table (3): Represents frequency table of degree variable

Degree	Frequency	Percent %
First Stage	9	16.98
Second Stage	9	16.98
Third Stage	10	18.87
Fourth Stage	25	47.17
Total	**53**	**100**

It is obvious from table (3) that the most of patients of this sample are reached to worst stage of cancer dieses which is fourth stage.

Table (4): Clarifies frequency table of occupation variable

Occupation	Frequency	Percent
Worker	1	1.89
Employee	8	15.09
Dealer	9	16.98
Teacher	3	5.66
House wife	19	35.85
Pensioner	6	11.32
Peshmarga	4	7.55
Builder	3	5.66
Total	**53**	**100**

From the above table it is clear that the most of patients are house wife followed by dealer and employee which are 35.85%, 16.98% and 15.09 respectively.

Second step: estimating the model, at beginning we should achieve the assumption of that the response variable does not followed a statistical distribution this can be done by using chi-square test as it is shown below for some common statistical distributions:

Table (5): Represents the test of response distribution

Distributions	Chi-Square	P-value
Exponential	107.92	0.0000
Gamma	112.52	0.0000
Gen.Gamma	136.77	0.0000
Lognormal	142.40	0.0000
Weibull	115.29	0.0000

From the above table it is obvious that the p-value is less than 0.05 which implies that the response has a free distribution.

DOI: http://dx.doi.org/10.25098/3.1.7

The Scientific Journal of Cihan University – Sulaimanyia **PP: 94-108**
Volume (3), Issue (1), Jun 2019
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

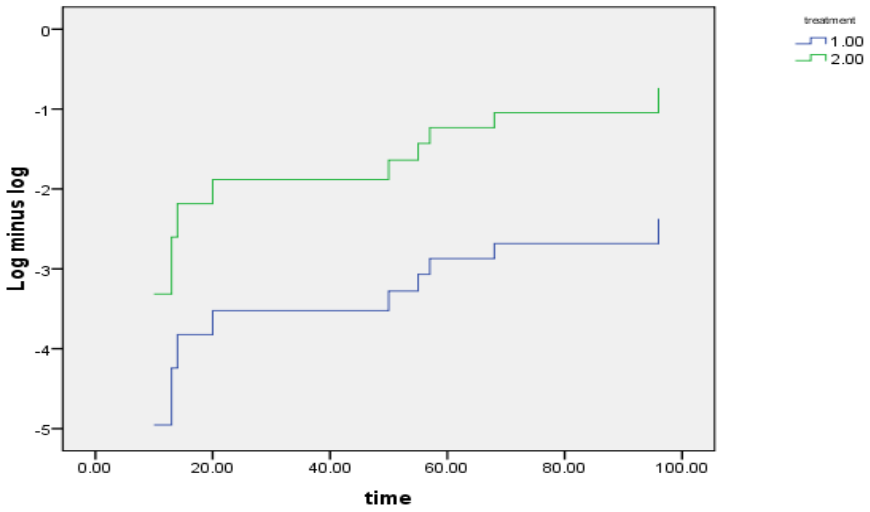

Figure (1): Represents the test of proportional hazard by using K-M

It is clear from the above figure that the proportional hazard is fixed then we achieved the assumption of Cox-Regression. Partial likelihood method has been used to estimate the model the results are showed up as follow:

Table(6): Represents the estimated parameters of the model

		Bi	SE	Wald	df	Sig.(p-value)
Step 1	Age	-0.053	0.06	0.757	1	0.384
	White blood cells(Wbc)	-0.037	0.061	0.361	1	0.548
	Hemoglobin(Hb)	-0.992	0.308	10.358	1	0.001
	Weight	0.162	0.082	3.925	1	0.048
	Number of Chemotherapy	-0.219	0.102	4.659	1	0.031
Step 2	Age	-0.049	0.062	0.602	1	0.438
	Hemoglobin(Hb)	-0.945	0.299	10.007	1	0.002
	Weight	0.161	0.082	3.889	1	0.049
	Number of Chemotherapy	-0.223	0.101	4.845	1	0.028
Step 3	Hemoglobin(Hb)	-0.872	0.29	9.059	1	0.003
	Weight	0.12	0.06	3.966	1	0.046
	Number of Chemotherapy	-0.245	0.101	5.815	1	0.016

Some to up table each of (Hemoglobin (Hb), Weight and number of chemotherapy) are statistically have a significant effect on survival time with coefficients (-0.872, 0.012, -0.245) respectively.

The Scientific Journal of Cihan University – Sulaimanyia **PP: 94-108**
Volume (3), Issue (1), Jun 2019
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

For selecting the best model backward method was used in three steps, the third step is best postulated model to predicting the survival time.

Table (7): Represents the Likelihood ratio test

Model selection	-2 Log Likelihood	Chi-square	d.f	Sig.(p-value)	Cox & Snell R^2
Model-1	50.84166302	22.6601	5	0.000392	0.21074
Model-2	51.19232857	21.0420	4	0.000311	0.20450
Model-3	51.79431038	19.4217	3	0.000224	0.19022

As it is shown from table (7) the third model is more reliable than the others two models to represents the case of study and its p-value is less than 0.05 that means it is significant, also it has maximum log likelihood.

Table(8): Shows the baseline cumulative hazard function for Cox model

Time	Baseline Cum Hazard	Survival	SE	Cum Hazard
11	2.7951	0.9957126	0.0052736	0.0042967
32	6.5858	0.9859123	0.0152409	0.0141879
80	16.1169	0.9671233	0.0255182	0.0334293
118	28.0123	0.9214478	0.0523064	0.0818092
120	42.537	0.8674560	0.0784429	0.1421905
229	60.2726	0.8058015	0.0972018	0.2159179
253	82.2589	0.7364282	0.1221550	0.3059435
289	110.784	0.6586611	0.1373401	0.4175461
294	150.8806	0.5698743	0.1644120	0.5623395
381	216.3688	0.4649293	0.1746787	0.7658699
383	301.6006	0.3334415	0.1845429	1.0982877
425	414.7263	0.2163356	0.1643399	1.5309244
651	577.3869	0.1218274	0.1204563	2.1051501
719	953.9579	0.0533535	0.0721667	2.9308153
767	2081.7629	0.0078890	0.0140961	4.8422895
795	6254.1515	0.0000258	0.0001545	10.5670264
840	.	0.0000000	0.0000000	31.7460669

DOI: http://dx.doi.org/10.25098/3.1.7

The Scientific Journal of Cihan University – Sulaimanyia **PP: 94-108**
Volume (3), Issue (1), Jun 2019
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Some to up table it is clear that the cumulative hazard is increased which leads to increasing the number of deaths for the patients under study.

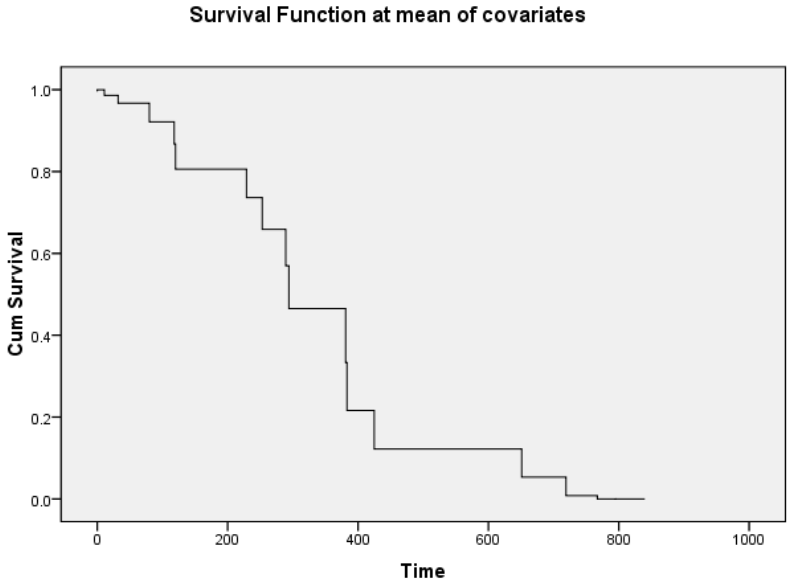

Figure (1): Represents the survival function

Some to up figure it is obvious that the survival time for patients is decreased. The survival time of the patients that spent 110 to 220 and 400 to 630 days receiving treatments their survival times are 0.80 and 0.12 respectively.

11. Conclusions:

From the results we can conclude that:

1. The data under study achieved the assumptions of using Cox-Regression.
2. Backward method gives three possible models of Cox where the third one is the more adequate model than the two other models.
3. According to the results in table (6) if the weight of the patient stays stable through taking treatments for period of time then the survival time will increase.
4. Increasing the number of chemo leads to decreasing the survival time of patient.
5. Through figure (2) the patients that spent 110 to 220 and 400 to 630 days receiving treatments their survival times are 0.80 and 0.12 respectively.
6. From table (8) one can conclude that the survival time of patient is decrease sharply because the cumulative hazard is increasing.

References:

[1] D. R. Cox *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 34, No. 2. (1972), pp.187-220.

[2] J. P. Klein and M. L. MoescHemoglobin(Hb)erger, *Survival Analysis: Techniques for Censored and Truncated Data*, Springer, New York, NY, USA, 2nd edition, 2003.

[3] Kleinbaum DG, Klein M, editors. Survival Analysis. 3rd ed. New York: Springer; 2012.

[4] Cox D.R., "partial likelihood", biometric , 62, 2 , p(269-276), 1975.

[5] Agresti A."an introduction to categorical data analysis". Wiley series in probability and ststistics, florida, 2007.

[6] Izenman , A.J. and Tran, L.T.,"Estimation of the survival function and hazard rate", Journal of stat planning and Inference, V .24,p(233-247), 1990.

[7] Ferlay J, Soerjomataram I, Ervik M, *et al. GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC Cancer Base No. 11*. 2014; Available from: globocan.iarc.fr

[8] Biglarian A, Hajizadeh E, Gouhari MR, Khodabakhshi R. Survival analysis of patients with gastric adenocarcinomas and factors related. Kowsar Med J. 2008;12:337–347.

[9] Zeraati H, Mahmoudi M, Kazemnejad A, Mohammad K. Postoperative survival in gastric cancer patients and its associated factors: a time dependent covariates model. Iranian J Public Health. 2006;35:40–46.

[10] Hisamichi S (1989) Screening for gastric cancer. World J Surg 13: 31-37.

[11] Ekman A. Variable selection for the Cox proportional hazards model : A simulation study comparing the stepwise, lasso and bootstrap approach [Internet] [Dissertation]. 2017. Available from: http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-130521

[12] Jessica Ko. Solving the Cox Proportional Hazards Model and Its Applications. May 20, 2017. EECS Department University of California, Berkele Technical Report No. UCB/EECS-2017-110

[13] Alhasawi, Eman. Survival analysis approaches for prostate cancer. 2015.

[14] Gellar, J. E., Colantuoni, E., Needham, D. M., & Crainiceanu, C. M. Cox regression models with functional covariates for survival data. (2015). *Statistical Modelling*, *15*(3), 256–278. https://doi.org/10.1177/1471082X14565526

[15] Pham, Minh Hoang .Survival Analysis - Breast Cancer. 2014. Vol. 6: Iss. 1, Article 4. DOI: http://dx.doi.org/10.5038/2326-3652.6.1.4860 .Available at: https://scholarcommons.usf.edu/ujmm/vol6/iss1/4

[16] Ahmadi, A., Roudbari, M., Gohari, M.R., & Hosseini, B. Estimation of hazard function and its associated factors in gastric cancer patients using wavelet and kernel smoothing methods. 2012. *Asian Pacific journal of cancer prevention : APJCP, 13 11*, 5643-6 .

[17] Cox, D. R., and E. J. Snell. *The Analysis of Binary Data*, 2nd ed. 1989. London: Chapman and Hall.

DOI: http://dx.doi.org/10.25098/3.1.7