Use Proportional Hazards Regression Method To Analyze The Survival of Patients with Cancer Stomach At A Hiwa Hospital /Sulaimaniyah

Mohammad M.Faqi Hussain *Faqi_zanko@yahoo.co.uk*
School of Admin&Eco / Statistic department
University of Sulaimani
Sulaimaniyah/Kurdistan/Iraq

Abstract

The Kaplan Meier method is used to analyze data based on the survival time. In this paper used Kaplan Meier procedure and Cox regression with these objectives. The objectives are finding the percentage of survival at any time of interest, comparing the survival time of two studied groups and examining the effect of continuous covariates with the relationship between an event and possible explanatory variables. The variables (Age, Gender, Weight, Drinking, Smoking, District, Employer, Blood Group) are used to study the survival patients with cancer stomach. The data in this study taken from Hiwa/Hospital in Sualamaniyah governorate during the period of (48) months starting from (1/1/2010) to (31/12/2013) .After Appling the Cox model and achieve the hypothesis we estimated the parameters of the model by using (Partial Likelihood) method and then test the variables by using (Wald test) the result show that the variables age and weight are influential at the survival of time.

Keywords: Survival Time, The Kaplan Meier Method, Cox Regression Method.

1. INTRODUCTION

This program performs Cox (proportional hazards) regression analysis, which models the relationship between a set of one or more covariates and the hazard rate. Covariates may be discrete or continuous. Cox's proportional hazards regression model is solved using the method of marginal likelihood outlined [8].

This routine can be used to study the impact of various factors on survival. You may be interested in the impact of diet, age, amount of exercise, and amount of sleep on the survival time after an individual has been diagnosed with a certain disease such as cancer. Under normal conditions, the obvious statistical tool to study the relationship between a response variable (survival time) and several explanatory variables would be multiple regression. Unfortunately, because of the special nature of survival data, multiple regressions is not appropriate. Survival data usually contain censored data and the distribution of survival times is often highly skewed. These two problems invalidate the use of multiple regressions. Many alternative regression methods have been suggested. The most popular method is the proportional hazard regression method.[4]

2. METHODOLOGIES

SPSS was used in this analysis. Kaplan Meier and Cox regression are the two main analyses in this paper. The Kaplan Meier procedure is used to analyze on censored and uncensored data for the survival time. It is also used to compare two treatment groups on their survival times. The Kaplan Meier technique is the univariate version of survival analysis. To present more details in the survival analysis, further analysis using Cox regression as multivariate analysis is presented. Cox regression allows the researcher to include predictor variables (covariates) into the models. Cox regression will handle the censored cases correctly. It will provide estimated coefficients for

each of the covariates that allow us to assess the impact of multiple covariates in the same model. We can also use Cox regression to examine the effect of continuous covariates. The steps required in SPSS to perform the above objectives are listed as follows.

3. THE COX REGRESSION MODEL [3][4]

Survival analysis refers to the analysis of elapsed time. The response variable is the time between a time origin and an end point. The end point is either the occurrence of the event of interest, referred to as a death or failure, or the end of the subject's participation in the study. These elapsed times have two properties that invalidate standard statistical techniques, such as t-tests, analysis of variance, and multiple regressions. First of all, the time values are often positively skewed. Standard statistical techniques require that the data be normally distributed. Although this skewness could be corrected with a transformation, it is easier to adopt a more realistic data distribution.

The second problem with survival data is that part of the data are censored. An observation is censored when the end point has not been reached when the subject is removed from study. This may be because the study ended before the subject's response occurred, or because the subject withdrew from active participation. This may be because the subject died for another reason, because the subject moved, or because the subject quit following the study protocol. All that is known is that the response of interest did not occur while the subject was being studied.

When analyzing survival data, two functions are of fundamental interest—the survivor function and the hazard function. Let T be the survival time. That is, T is the elapsed time from the beginning point, such as diagnosis of cancer, and death due to that disease. The values of T can be thought of as having a probability distribution.

Suppose the *probability density function* of the random variable *T* is given by $f(T)$. The *probability distribution function* of *T* is then given by

$$F(T) = \Pr(t < T) = \int_0^T f(t)dt \qquad (1)$$

The *survivor function*, $S(T)$, is the probability that an individual survives past *T*. This leads to

$$S(T) = \Pr(T > t) = 1 - F(t) \qquad (2)$$

The *hazard function* is the probability that a subject experiences the event of interest (death, relapse, etc.) during a small time interval given that the individual has survived up to the beginning of that interval. The mathematical expression for the hazard function is

$$h(T) = \lim_{\Delta T \to o} \frac{\Pr(T \le t < (T + \Delta T) | T \le t)}{\Delta T} = \lim_{\Delta T \to o} \frac{F(T + \Delta T) - F(T)}{\Delta T} = \frac{f(T)}{S(T)} \qquad (3)$$

The cumulative hazard function $H(T)$ is the sum of the individual hazard rates from time zero to time *T*. The formula for the cumulative hazard function is

$$H(T) = \int_0^T h(u)\,du$$

Thus, the hazard function is the derivative, or slope, of the cumulative hazard function. The cumulative hazard function is related to the cumulative survival function by the expression

$$S(T) = e^{-H(T)} \qquad (4)$$

Or

$$H(T) = -\ln(S(T)) \qquad (5)$$

We see that the distribution function, the hazard function, and the survival function are mathematically related. As a matter of convenience and practicality, the hazard function is used in the basic regression model.

Cox (1972) expressed the relationship between the hazard rate and a set of covariates using the model

$$\ln[h(T)] = \ln[h_0(T)] + \sum_{i=1}^{p} x_i \beta_i$$

or

$$h(T) = h_0(T) e^{\sum_{i=1}^{p} x_i \beta_i} \qquad (6)$$

Where $x_1, x_2, ..., x_p$ are covariates, $\beta_1, \beta_2, ..., \beta_p$ are regression coefficients to be estimated, *T* is the elapsed time, and $h_o(T)$ is the baseline hazard rate when all covariates are equal to zero. Thus the linear form of the regression model is

$$\ln[\frac{h(T)}{h_0(T)}] = \sum_{i=1}^{p} x_i \beta_i$$

Taking the exponential of both sides of the above equation, we see that this is the ratio between the actual hazard rate and the baseline hazard rate, sometimes called the *relative risk*. This can be rearranged to give the model

$$\frac{h(T)}{h_0(T)} = \exp(\sum_{i=1}^{p} x_i \beta_i) = e^{x_1 \beta_1} e^{x_2 \beta_2} ... e^{x_p \beta_p} \qquad (7)$$

The regression coefficients can thus be interpreted as the relative risk when the value of the covariate is increased by one unit.

Note that unlike most regression models, this model does not include an intercept term. This is because if an intercept term were included, it would become part of $h_o(T)$.

Also note that the above model does not include *T* on the right-hand side. That is, the relative risk is constant for all time values. This is why the method is called *proportional hazards*.

An interesting attribute of this model is that you only need to use the ranks of the failure times to estimate the regression coefficients. The actual failure times are not used except to generate the ranks. Thus, you will achieve the same regression coefficient estimates regardless of whether you enter the time values in days, months, or years.

3.1 Cumulative Hazard [1][6]

Under the proportional hazards regression model, the cumulative hazard is

$$H(T,X)=\int_0^T h(u,X)\,du=\int_0^T h_o(u)\,e^{\sum_{i=1}^{p}x_i\beta_i}\,du=e^{\sum_{i=1}^{p}x_i\beta_i}\int_0^T h_o(u)\,du$$
$$=H_o(T)e^{\sum_{i=1}^{p}x_i\beta_i} \quad (8)$$

Note that the survival time *T* is present in $H_o(T)$, but not in $e^{\sum_{i=1}^{p}x_i\beta_i}$. Hence, the cumulative hazard up to time *T* is represented in this model by a baseline cumulative hazard $H_o(T)$ which is adjusted by the covariates by multiplying by the factor $e^{\sum_{i=1}^{p}x_i\beta_i}$.

3.2 Cumulative Survival [1][6]

Under the proportional hazards regression model, the cumulative survival is

$$S(T,X)=\exp(-H(T,X))=\exp(-H_o(T)e^{\sum_{i=1}^{p}x_i\beta_i})=[e^{-H_o(T)}]^{e^{\sum_{i=1}^{p}x_i\beta_i}}$$
$$=S_o(T)^{e^{\sum_{i=1}^{p}x_i\beta_i}} \quad (9)$$

Note that the survival time *T* is present in $S_o(T)$, but not in $e^{\sum_{i=1}^{p}x_i\beta_i}$

3.3 Maximum Likelihood Estimation [1][6]

Let $t=1,..,M$ index the *M* unique failure times $T_1,T_2,...,T_M$. Note that *M* does not include duplicate times or censored observations. The set of all failures (deaths) that occur at time T_t is referred to as D_t . Let index the members of . The set of all individuals that are at risk immediately before time T_t is referred to as R_t . This set, often called the *risk set*, includes all individuals that fail at time as well as those that are censored or fail at a time later than T_t . Let $r=1,...,n_t$ index the members of R_t . Let *X* refer to a set of *p* covariates. These covariates are indexed by the subscripts *i, j,* or *k*. The values of the covariates at a particular failure time T_d are written $x_{1d},x_{2d},...,x_{pd}$ or x_{id} in general. The regression coefficients to be estimated are $\beta_1,\beta_2,...,\beta_p$

3.4 The Log Likelihood

When there are no ties among the failure times, the log likelihood is given as follows [8]:

$$LL(\beta)=\sum_{i=1}^{M}\{(\sum_{i=1}^{p}x_{it}\beta_i)-\ln(\sum_{r\in R_t}\exp(\sum_{i=1}^{p}x_{it}\beta_i))\}$$
$$=\sum_{i=1}^{M}\{(\sum_{i=1}^{p}x_{it}\beta_i)-\ln(G_{R_t})\} \quad (10)$$

Where $$G_R = (\sum_{r \in R_t} \exp(\sum_{i=1}^{p} x_{it}\beta_i)$$

The following notation for the first-order and second-order partial derivatives will be useful in the derivations in this section.

$$\begin{aligned} H_{jR} &= \frac{\partial G_R}{\partial \beta_j} = \sum_{r \in R} x_{jr} \exp(\sum_{i=1}^{p} x_{it}\beta_i) \\ A_{jkR} &= \frac{\partial^2 G_R}{\partial \beta_j \partial \beta_k} = \frac{\partial H_{jR}}{\partial \beta_k} = \sum_{r \in R} x_{jr} x_{kr} \exp(\sum_{i=1}^{p} x_{it}\beta_i) \end{aligned} \quad (11)$$

The maximum likelihood solution is found by the Newton-Raphson method. This method requires the first and second order partial derivatives. The first order partial derivatives are

$$U_j = \frac{\partial LL(\beta)}{\partial \beta_j} = \sum_{i=1}^{M} \{x_{jt} - \frac{H_{jR_t}}{G_{R_t}}\} \quad (12)$$

The second order partial derivatives, which are the information matrix, are

$$I_{jk} = \sum_{i=1}^{M} \frac{1}{G_{R_t}} (A_{jkR_t} - \frac{H_{jR_t}}{G_{R_t}}) \quad (13)$$

When there are failure time ties (note that censor ties are not a problem), the exact likelihood is very cumbersome [2][7]. Breslow's approximation was used by the first Cox regression programs, but Efron's approximation provides results that are usually closer to the results given by the exact algorithm and it is now the preferred approximation [6]. We have included Breslow's method because of its popularity.

3.5 Breslow's Approximation To The Log Likelihood

The log likelihood of Breslow's approximation is given as follows:- [8]

$$\begin{aligned} LL(\beta) &= \sum_{i=1}^{M} \{(\sum_{d \in D_t} \sum_{i=1}^{p} x_{id}\beta_i) - m_t \ln(\sum_{r \in R_t} \exp(\sum_{i=1}^{p} x_{it}\beta_i))\} \\ &= \sum_{i=1}^{M} \{(\sum_{d \in D_t} \sum_{i=1}^{p} x_{id}\beta_i) - m_t \ln(G_{R_t})\} \end{aligned} \quad (14)$$

Where

$$G_R = (\sum_{r \in R_t} \exp(\sum_{i=1}^{p} x_{it}\beta_i)$$

The maximum likelihood solution is found by the Newton-Raphson method. This method requires the first-order and second-order partial derivatives. The first order partial derivatives are

$$\begin{aligned} U_j &= \frac{\partial LL(\beta)}{\partial \beta_j} \\ &= \sum_{i=1}^{M} \{(\sum_{d \in D_t} x_{jd}) - (m_t \frac{H_{jR_t}}{G_{R_t}})\} \end{aligned} \quad (15)$$

The negative of the second-order partial derivatives, which form the information matrix, are

$$I_{jk} = \sum_{i=1}^{M} \frac{m_t}{G_{R_t}} (A_{ijkR_t} - \frac{H_{jR_t} H_{kR_t}}{G_{R_t}}) = \sum_{i=1}^{M} \{ (\sum_{d \in D_t} x_{jd}) - (m_t \frac{H_{jR_t}}{G_{R_t}}) \} \quad (16)$$

3.6 Efron's Approximation to the Log Likelihood

The log likelihood of Efron's approximation is given as follows [8]:-

$$LL(\beta) = \sum_{t=1}^{M} \{ (\sum_{d \in D_t} \sum_{i=1}^{p} x_{id} \beta_i) - \sum_{d \in D_t} \ln[\sum_{r \in R_t} \exp(\sum_{i=1}^{p} x_{it} \beta_i) - \frac{d-1}{m_t} \sum_{c \in R_t} \exp(\sum_{i=1}^{p} x_{ic} \beta_i) \} \quad (17)$$
$$= \sum_{t=1}^{M} \{ \sum_{d \in D_t} \sum_{i=1}^{p} x_{id} \beta_i - \sum_{d \in D_t} \ln[G_{R_t} - \frac{d-1}{m_t} G_{D_t}] \}$$

The maximum likelihood solution is found by the Newton-Raphson method. This method requires the first and second order partial derivatives. The first partial derivatives are

$$U_j = \frac{\partial LL(\beta)}{\partial \beta_j} = \sum_{t=1}^{M} \{ (\sum_{d \in D_t} x_{jd}) - (m_t \frac{H_{jR_t} - (\frac{d-1}{m_t}) H_{jD_t}}{G_{R_t} - (\frac{d-1}{m_t}) G_{D_t}}) \} \quad (18)$$
$$= \sum_{t=1}^{M} \sum_{d \in D_t} x_{jd} - \sum_{t=1}^{M} \sum_{d \in D_t} (\frac{H_{jR_t} - (\frac{d-1}{m_t}) H_{jD_t}}{G_{R_t} - (\frac{d-1}{m_t}) G_{D_t}})$$

The second partial derivatives provide the information matrix which estimates the covariance matrix of the estimated regression coefficients. The negative of the second partial derivatives are

$$U_j = -\frac{\partial^2 LL(\beta)}{\partial \beta_j \partial \beta_k} \quad (19)$$
$$= \sum_{t=1}^{M} \sum_{d=1}^{m_t} \frac{[G_{R_t} - (\frac{d-1}{m_t}) G_{D_t}][A_{jkR_t} - (\frac{d-1}{m_t}) A_{jkD_t}][H_{jR_t} - (\frac{d-1}{m_t}) H_{jD_t}][H_{kR_t} - (\frac{d-1}{m_t}) H_{kD_t}]}{[G_{R_t} - (\frac{d-1}{m_t}) G_{D_t}]^2}$$

4. ESTIMATION OF THE SURVIVAL FFUNCTION

Once the maximum likelihood estimates have been obtained, it may be of interest to estimate the survival probability of a new or existing individual with specific covariate settings at a particular point in time [8][12].

4.1 Cumulative Survival

This estimates the cumulative survival of an individual with a set of covariates all equal to zero. The survival for an individual with covariate values of is X_0

$$S(T, X_o) = \exp(-H(T \mid X_o)) = \exp(H_o(T \mid X_o) \exp \sum_{i=1}^{p} x_{io}\beta\ ) \qquad (20)$$
$$= S_o(T)^{\exp \sum_{i=1}^{p} x_{io}\beta}$$

T.0h23e estimate of the baseline survival function is calculated from the cumulated hazard function using $S_o(T)$.

$$S_o(T) = \prod_{T_t \le T_o} \alpha_t$$

Where

$$\alpha_t = \frac{S(T_t)}{S(T_{t-1})} = [\frac{S_o(T_t)}{S_o(T_{t-1})}]^{\exp(\sum_{i=1}^{p} x_{it}\beta_i)} = [\frac{S_o(T_t)}{S_o(T_{t-1})}]^{\theta_t} \quad , \ \theta_r = \exp \sum_{i=1}^{p} x_{ir}\beta_i \qquad (21)$$

The value of α*t*, the conditional baseline survival probability at time *T*, is the solution to the conditional likelihood equation

$$\alpha_t = \sum_{d \in D_t} \frac{\theta_d}{1 - \alpha_t^{\theta_d}} = \sum_{r \in R_t} \theta_r$$

When there are no ties at a particular time point *Dt*, contains one individual and the above equation can be solved directly, resulting in the solution

$$\hat{\alpha}_t = [1 - \frac{\hat{\theta}_t}{\sum_{r \in R_t} \hat{\theta}_r}]^{\hat{\theta}_t - 1} \qquad (22)$$

When there are ties, the equation must be solved iteratively. The starting value of this iterative process is

$$\hat{\alpha}_t = \exp[\frac{-m_t}{\sum_{r \in R_t} \hat{\theta}_r}]^{\hat{\theta}_t - 1} \qquad (23)$$

4.2 baseline hazard Rate [6][12][13]

Estimate the baseline hazard rate as follows $h_o(T_t)$

$$h_o(T_t) = 1 - \alpha_t \qquad (24)$$

They mention that this estimator will typically be too unstable to be of much use. To overcome this, you might smooth these quantities using lowess function of the Scatter Plot program.

4.3 Cumulative Hazard [6][12][13]

An estimate of the cumulative hazard function $H_o(T)$ derived from relationship between the cumulative hazard and the cumulative survival. The estimated baseline survival is

$$\hat{H}_o(T) = -\ln(\hat{S}_o(T))$$

This leads to the estimated cumulative hazard function is

$$\hat{H}(T) = -\exp(\sum_{i=1}^{p} x_i \hat{\beta}_i) \ln(\hat{S}_o(T)) \qquad (25)$$

4.4 Cumulative Survival [6][12][13]
The estimate of the cumulative survival of an individual with a set of covariates values of X_0 is

$$\hat{S}_o(T \mid X_o)\hat{H}(T) = \hat{S}_o(T)^{\exp(\sum_{i=1}^{p} x_{io}\hat{\beta}_i)} \qquad (26)$$

5. STATISTICAL TEST AND CONFIDENCE INTERVAL [6][12][13]

Inferences about one or more regression coefficients are all of interest. These inference procedures can be treated by considering hypothesis tests and/or confidence intervals. The inference procedures in Cox regression rely on large sample sizes for accuracy. Two tests are available for testing the significance of one or more independent variables in a regression: the likelihood ratio test and the Wald test. Simulation studies usually show that the likelihood ratio test performs better than the Wald test. However, the Wald test is still used to test the significance of individual regression coefficients because of its ease of calculation. These two testing procedures will be described next.

6. LIKLIHOOD RATION AND DEVIANCE [3][4][6]

The *Likelihood Ratio* test statistic is -2 times the difference between the log likelihoods of two models, one of which is a subset of the other. The distribution of the LR statistic is closely approximated by the chi-square distribution for large sample sizes. The degrees of freedom (DF) of the approximating chi-square distribution is equal to the difference in the number of regression coefficients in the two models. 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. That is, if *L*full is the log likelihood of the full model and *L*subset is the log likelihood of a subset of the full model, the likelihood ratio is defined as

$$LR = -2[L_{subset} - \mathrm{L}_{\mathrm{full}}] = -2[\ln(\frac{l_{subset}}{l_{full}})] \qquad (27)$$

Note that the -2 adjusts *LR* so the chi-square distribution can be used to approximate its distribution. The likelihood ratio test is the test of choice in Cox regression. Various simulation studies have shown that it is more accurate than the Wald test in situations with small to moderate sample sizes. In large samples, it performs about the same. Unfortunately, the likelihood ratio test requires more calculations than the Wald test, since it requires the fitting of two maximum-likelihood models.

7. Deviance [3][4][6]

When the full model in the likelihood ratio test statistic is the saturated model, *LR* is referred to as the *deviance*. A saturated model is one which includes all possible terms (including interactions) so that the predicted values from the model equal the original data. The formula for the deviance is

$$D = -2[L_{\mathrm{Re}\,duced} - \mathrm{L}_{\mathrm{Saturated}}] \qquad (28)$$

The deviance in Cox regression is analogous to the residual sum of squares in multiple regression. In fact, when the deviance is calculated in multiple regression, it is equal to the sum of the squared residuals.

The change in deviance, Δ*D*, due to excluding (or including) one or more variables is used in Cox regression just as the partial *F* test is used in multiple regression. Many texts use the letter *G* to represent Δ*D*. Instead of using the *F* distribution, the distribution of the change in deviance is approximated by the chi-square distribution. Note that since the log likelihood for the saturated model is common to both deviance values , Δ*D* can be calculated without actually fitting the saturated model. This fact becomes very important during subset selection. The formula for Δ*D* for testing the significance of the regression coefficient(s) associated with the independent variable X_1 is

$$\begin{aligned}\Delta D_{X1} &= D_{without\,X1} - D_{with\,X1} = -2[\,L_{withoutX1} - L_{Saturated}\,] + 2[L_{withX1} - L_{Saturated}\,] \\ &= -2[\,L_{without\,X1} - D_{with\,X1}\,]\end{aligned} \qquad (29)$$

Note that this formula looks identical to the likelihood ratio statistic. Because of the similarity between the change in deviance test and the likelihood ratio test, their names are often used interchangeably.

7.1 Wald test [3][4][6]

The Wald test will be familiar to those who use multiple regression. In multiple regression, the common *t*-test for testing the significance of a particular regression coefficient is a Wald test. In Cox regression, the Wald test is calculated in the same manner. The formula for the Wald statistic is

$$Z_j = \frac{b_j}{S_{b_j}} \qquad (30)$$

Where S_{b_j} is an estimate of the standard error of b_j provided by the square root of the corresponding diagonal element of the covariance matrix, $Var(\hat{\beta}) = I^{-1}$.

With large sample sizes z_j, the distribution is closely approximated by the normal distribution. With small and moderate sample sizes, the normal approximation is described as ‘adequate.

7.2 Confidence Intervals [3][4][6]

Confidence intervals for the regression coefficients are based on the Wald statistics. The formula for the limits of a two-sided confidence interval is 100 (1−α)%

$$b_j \pm |\, z_{\frac{\alpha}{2}} \,|\, S_{b_j} \qquad (31)$$

7.3 Coefficient of determination R2 [6]

The time of the writing of their book, there is no single, easy to interpret measure in Cox regression that is analogous to R^2 in multiple regression. They indicate that if such a measure “must be calculated” they would use

$$R_p^2 = 1 - \exp[\frac{2}{n}(L_o - L_p)] \qquad (32)$$

Where L_0 is the log likelihood of the model with no covariates, *n* is the number of observations (censored or not), and *Lp* is the log likelihood of the model that includes the covariates.

8. DESCRIPTION DATA

In this paper, depending on real data for the cancer stomach diseases , the researcher choosing this type of cancer because it is diffusion and in the current time in Sulaimaniyah governorate / Iraq Kurdistan region . To collect the data for the cancer stomach diseases, returning the atomic medicine and radiance in the hiwa hospital in Sulaimaniyah. The time of the collecting data for this study between (1/1/2010-31/12/2013).

9. RESEARACH VARIABLES

The most important variables that have been studied in this paper

T: Survival Time [is the survival time of patients (cancer stomach) until death or Surveillance]
D : Represent variable case [1 :Death 0: Censored]
X_1: Represents the gender of the patient [1: Male 2:Female]
X_2: Repersents the patient's age at injury
X_3: Represents the patient's weight at injury
X_4: Represents the Smoking Variable [1: Smoking 2:Non-Smoking]
X_5 : Represents the Blood group Variable
X_6 : Represents the district Variable [1: Within the governorate 2: Outside governorate]
X_7 : Represents the Drinking Variable [1: Drinking 2:Non-Drinking]

10. USING THE KAPLAN MEIER PROCEDURE

Table 1, Table 2, and Figure 1 are presented as below, the analysis based on the first objective. Table 1 shows the number of events, namely the number of cases is 5, with the percentage of censored cases being 97.6%. Table 2 shows the mean of survival time is 242 months, with the standard error 26.756. Figure 1 shows the survival plots. It is shown in the diagram that at 300 months, 82% of the observations were still alive. The Figure 1, show more information on the percentage of the survival for different months can be accessed by referring to the specific month and looking for the associate survival rate.

Total N	N of Events	Censored	
		N	Percent
210	5	205	97.6%

TABLE 1: Case Processing Summary.

Mean Estimate	Std. Error	95% Confidence Interval	
		Lower Bound	Upper Bound
242.046	26.756	189.605	294.487

a. Estimation is limited to the largest survival time if it is censored.

TABEL 2: Means for Survival Time.

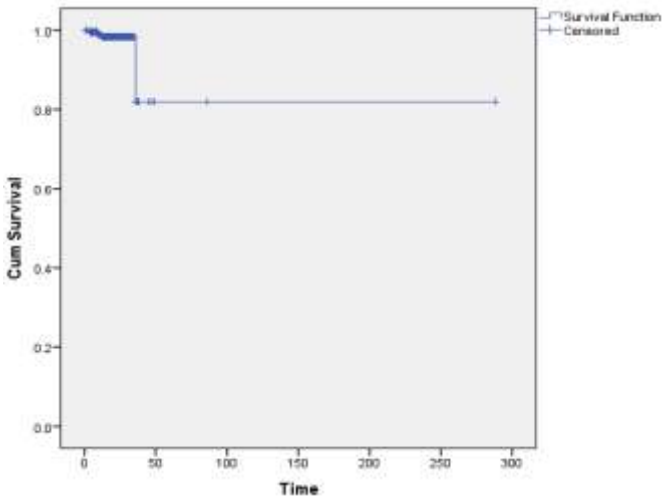

FIGURE 1: Survival Function.

Table 3, Table 4, Table 5 and Figure 2 are presented as below, the analysis of the second objective. Table 3 shows the number of cases for the two categories in Age, with cases of 1-40 years (164 observations) and cases 41-80 year (46 observations). Thus, there are 164 observations which have cancer stomach in the range (1-40) years and 46 observations which have cancer stomach in the range (41-80) year. Table 4 shows the mean survival times for the two groups, with the mean for cases of (1-40) years is 79.966 months and the mean for cases (41-80) years is 193.230 months. Table 5 shows the results of log-rank test with the p-value of 0.043, which indicates that there is a significant difference between the two groups having a shorter time to event. The survival plot (Figure 2) shows that the group (41-80) years has a longer survival time to event compared to the group (1-40) years. This situation is shown in Figure 2, where 62% of patients with range (1-40) year was still alive at 48 months as compared with 84% of patients with range (41-80) years. From Figure 2, more information about the survival rate for different months for the two groups can be retrieved by referring to the specific month and looking for the associated survival rates.

Age	Total N	N of Events	Censored	
			N	Percent
1-40	164	2	162	98.8%
41-80	46	3	43	93.5%
Overall	210	5	205	97.6%

TABLE 3: Case Processing Summary.

Age	Mean Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1-40	79.966	5.227	69.721	90.212
41-80	193.230	65.180	65.476	320.983

TABLE 4: Means for Survival Time.

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	4.101	1	.043

TABLE 5: Overall Comparisons.

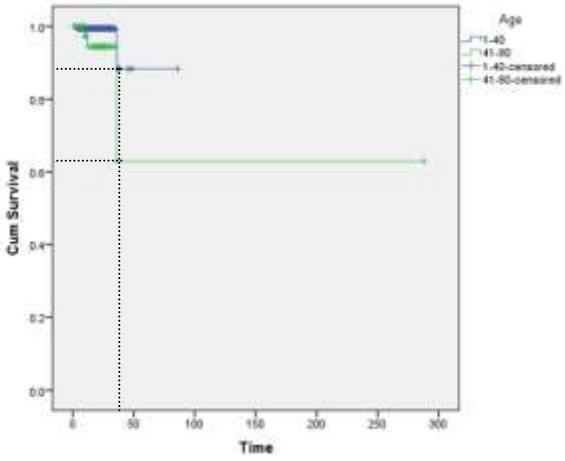

FIGURE 2: Survival Plot for Comparison of the Two Groups.

The results from the Cox regression are presented in Table 6, Table 7, Table 8, and Table 9. Table 6 shows that only 98.6% of the observations or cases are available in the analysis and there is no number of cases dropped.

Details		**N**	**Percent**
Cases available in analysis	Event	5	2.4%
	Censored	202	96.2%
	Total	207	98.6%
Cases dropped	Cases with missing values	0	0.0%
	Cases with negative time	0	0.0%
	Censored cases before the earliest event in a stratum	3	1.4%
	Total	3	1.4%
Total			210

TABLE 6: Case Processing Summary.

variables		Frequency	(1)	(2)	(3)	(4)	(5)	(6)
Gender	1=Male	127	1					
	2=Female	83	0					
Age	1=10-20	7	1	0	0	0		
	2=21-30	55	0	1	0	0		
	3=31-40	102	0	0	1	0		
	4=41-50	45	0	0	0	1		
	7=71-80	1	0	0	0	0		
Weight	1=31-50	39	1	0	0			
	2=51-70	155	0	1	0			
	3=71-90	13	0	0	1			
	4=91-110	3	0	0	0			
Employer	1=Worker	70	1	0	0	0	0	
	2=Housewife	83	0	1	0	0	0	
	3=Officer	20	0	0	1	0	0	
	4=Student	14	0	0	0	1	0	
	5=Child	1	0	0	0	0	1	
	6=Retard	22	0	0	0	0	0	
Smoking	1=Yes	88	1					
	2=No	122	0					
Blood _Group	1=A+	45	1	0	0	0	0	0
	2=A-	4	0	1	0	0	0	0
	3=B+	15	0	0	1	0	0	0
	4=B-	2	0	0	0	1	0	0
	5=AB+	55	0	0	0	0	1	0
	7=O+	82	0	0	0	0	0	1
	8=O-	7	0	0	0	0	0	0
District	1=Inside	118	1					
	2=Outside	92	0					
Drinking	1=Yes	12	1					
	2=No	198	0					

TABLE 7: Categorical Variable Coding a,c,d,e,f,g,h,i.

Table 8 shows the model is significant using chi-square test, the value of the test is equal to 17.587 and the p-value of the test is equal to (p-value=0.025) which is less than (0.05). Table 9 provides the p-values and the hazard ratio (Exp(B)) of the variables.SE values in Table 9 are

small, and the problem of multicolinearity is under controlled. For the confounder model, the most important variables to be looked into the group factors, which are the Age and Weight. The result shown that the p-value of age and weight are equal to 0.038 and 0.009 respectively, which are significant as reported in the Kaplan Meier analysis. The associate hazard ratio (HR) as indicated in Exp(B) is 0.05, which is less than ' 1'. For reporting HR, there are three possibilities: (a) a value of' 1' means there is no differences between two groups in having a shorter time to event, (b) a value of 'more than 1' means that the group of interest is likely to have a shorter time to event as compared to the reference group, and (c) a value of 'less than 1' means that the groups of interest less likely to have a shorter time to event comparing to the reference group. Therefore, the group of interest for Age (which is '1' –1-40 years) is less likely to have a shorter time to event (death) as compared to the reference group. Table 9 also shows that the Age and weight are significant, whereas other variables have insignificant.

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
23.568	17.587	8	.025	17.529	8	.025	17.529	8	.025

a. Beginning Block Number 1. Method = Enter

TABLE 8: Omnibus Tests of Model Coefficients.

Variables	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Age	3.395	1.640	4.287	1	.038	29.807	1.199	741.245
Weight	2.553	.973	6.881	1	.009	12.850	1.907	86.592
Drinking	-.339	2.875	.014	1	.906	.712	.003	199.557
Gender	2.337	1.660	1.982	1	.159	10.353	.400	268.062
Employer	.061	.560	.012	1	.913	1.063	.355	3.189
District	-1.317	1.359	.939	1	.332	.268	.019	3.844
Smoking	.721	1.546	.217	1	.641	2.055	.099	42.532
Blood Group	-.043	.307	.019	1	.890	.958	.525	1.751

TABLE 9: Variables in the Equation.

11. CONCLUSION AND RECOMANDATION

11.1 Conclusion

i) In the analysis part , the result show that the variables age and weight are two variables that effected to the survival time and choose these variables to stay in the model and other variables (drinks, sex, Employer, District, Smoking, Blood Group) has no significance effect to the survival time.

ii) In the analysis part, the biggest risk to be effect on survival time during (24-48) months, where the hazard rate is equal to (0.002314815) as shown in the table 10.

iii) The possibility of the survival times of patients decrease in the first period to the second period and fixed the survival times until reach the (288) months.

Interval Start Time	Number Entering Interval	Number Withdrawing during Interval	Number Exposed to Risk	Proportion Terminating	Cumulative Proportion Surviving at End of Interval	Hazard Rate	Std. Error of Hazard Rate
0-24	210	138	141	0.0212766	0.978723404	0.000896057	0.000517
24-48	69	64	37	0.05405405	0.925819436	0.002314815	0.001636
48-72	3	1	2.5	0	0.925819436	0	0
72-96	2	1	1.5	0	0.925819436	0	0
96-120	1	0	1	0	0.925819436	0	0
120-144	1	0	1	0	0.925819436	0	0
144-168	1	0	1	0	0.925819436	0	0
168-192	1	0	1	0	0.925819436	0	0
192-216	1	0	1	0	0.925819436	0	0
216-240	1	0	1	0	0.925819436	0	0
240-264	1	0	1	0	0.925819436	0	0
264-288	1	0	1	0	0.925819436	0	0
288-312	1	1	0.5	0	0.925819436	0	0

TABLE 10: Life Table.

11.2 Recommendations

i) Use cox model to conduct and more studies about the other types of cancers and knowing the factors that affecting at each type of these disease.

ii) The data is the primary factor in any study so we recommend the development of statistical cadres specialized in the field of organization and data arrangement in hospitals and health centers to register correctly.

12. REFRENCES

[1] Agresti A ”Categorical Data Analysis”, John Wiley and Sons, New York,1999.

[2] Breslow,N.E ”Covariance analysis of censored survival data”, Biometrics, 30, 89-100,1974.

[3] Collet, D. ”Modeling survival data in medical research”, Boca Raton, FL: Chapman & Hall/CRC,2003

[4] Cox, D.R.” Regression Models and Life tables (with discussion)”, Journal of the Royal Statistical Society, 34: 187—220,1972.

[5] Daniel, W. W. ” Biostatistics: A foundation for analysis in the health sciences”, River Street, U.S.: John Wiley & Sons, Inc.,2005.

[6] DW Hosmer, Jr., S Lemeshow ” Applied Survival Analysis: Regression Modeling of Time to Event Data”, New York: John Wiley, pp.386.1999.

[7] Efron, B.” The efficiency of Cox’s likelihood function for censored data”, Journal of the American Statistical Association 72, 557–565,1977.

[8] J.D. Kalbfleisch and R.L. Prentice ”The statistical analysis of failure time data”, John Wiley & Sons, Inc., New York,1980.

[9] Kaplan, E. L., & Meier, P.” Nonparametric estimation from incomplete Observations”, Journal of the American Statistical Association, 53, 45-81,1958.

[10] Komarek, A., Lesaffre, E., Harkanen, T., Declerck, D., & Virtanen, J. I. A Bayesian analysis of multivariate doubly-interval-censored dental data”, Biostatistics, 6(1), pp 145-155, 2005.

[11] Lesaffre, M. ”An overview of methods for interval-censored data with an emphasis on application in dentistry”, Statistical Methods in Medical Research,14(6), 539-552, 2005.

[12] Perrigot, R., Cliquet, G., & Mesbah, M. “ Possible applications of survival analysis in franchising research”, The International Review of Retail, Distribution and Consumer Research,14(1), 129-143, 2004.

[13] Walter A. Shewhart and Samuel S. Wilks ,”Weibull Models”, Johan Wiley & Sons. New York, 2004.