Baghdad Science Journal **Vol.6(3)2009**

Cox proportion hazard model for patients with hepatitis disease in Iraq

*Fadhaa O. Sameer**

Date of acceptance 3/3 / 2009

Abstract
Cox regression model have been used to estimate proportion hazard model for patients with hepatitis disease recorded in Gastrointestinal and Hepatic diseases Hospital in Iraq for (2002 -2005). Data consists of (age, gender, survival time terminal stat). A Kaplan-Meier method has been applied to estimate survival function and hazerd function.

Key words: Cox regression model, Kaplan-Meier method, Proportion hazard model, Survivor function, failure rate.

Introduction:
Survival analysis is a well known and widely used statistical procedure correlating time to event data. An event might be death, developing a certain disease, or any other condition if its occurrence can be clearly detected on time scale. Applying this method it is possible to present the probability [1].

The term survival data refers to the length of time, t, that corresponds to the time period from a well-defined start time t_0 until the occurrence of some particular event or end-point t_c, i.e. $t = t_c - t_0$. It is a common outcome measure in medical studies for relating treatment effects to the survival time of the patients. In these cases, the typical start time is when the patient inters the hospital, and the end point is when the patient died or was lost to follow-up[2]. In the follow-up process, not every individual ends up having the event of interest observed. Some have left the study before the failure occurred, or were simply lost in the follow-up, or the study closed. Thus, their true failure time should be longer than the observed In practice, survivals data are often collected from a large clinical trial are involved .In general; survival data have two distinctive features: non-symmetrical distributions and frequently censored observations [3]. The frequency plot for most survival data shows a longer 'tail' to the right (known as positive skew) that would not meet the assumption of Normality and survival data are termed right censored survival times and we make the assumption that the censoring event is independent of the true survival time. There are also cases of left-censored and interval censored data that will not be covered in this introduction [4].

Questions for survival data analysis
In substantive fields where a 'treatment' (e.g. a drug or surgery) may be introduced and evaluated in comparison to a control group, the main research questions can be summarized as follows.
1-How long on average are the subjects going to survive after the treatment?
2- Does a particular treatment result in a longer survival of subjects than other treatments?
3-What are the risk factors that may affect the survival time?
In this study, no particular medical treatment is involved. The general term 'survival time' means the survival days

*Tropical-Biological Researches Unit – college of science-Baghdad University

in hospital until death or living (censored). We shall be estimating the proportion hazard rate and survival function by using Cox regression model in order to assess their health well being.

Materials and Methods

1-Theorical Part

1-1 Survivor function [4]:

The probability that the random survival time variable T is greater than or equal to a specific t. Assuming F (t) is the cumulative Distribution function of t, the survivor function is the right tail probability, and So is defined as

$S(t) = P(T \geq t) = 1 - F(t)$ …. (1)

Where S(t) is survivor function.

1-2Hazard function [5]:

The probability that an individual dies at or just after time t, Conditional that having survived to that time. It represents the Instantaneous death rate for an individual surviving to time t or the probability that a case will terminate at time (t), and is defined as

$$h(t) = P(t \leq T < t + \Delta t) / p(T \geq t)$$
$$= [F(t + \Delta t) - F(t)] / S(t) \ldots (2)$$
$$(\Delta t \rightarrow 0)$$

The term Δt represents a very small unit increment of time.

1-3 Cumulative hazard function: The cumulative sum of the hazard probability Function that can be expressed as,

$H(t) = -\log S(t)$ … (3)

1-4 Median survival time:

The time when S (t) = 0.5. This statistic is termed the life expectancy in the population see.

1-5 Comparison of mean survival time or survivor function between groups, by means of statistical tests such as Log-rank taking into account the stratification in the data[6].

1-6Regression analysis for multiple explanatory variables associated with the median survival time or survival function or hazard function, by means of parametric survival models and semi-parametric proportional hazard models there are many well established statistical methods for carrying out these analyses. These are listed under the categories of non-parametric, parametric, and semi-parametric approaches see[2] .

Kaplan-Meier Method

Kaplan-Meier estimate of survivor and hazard functions Given n individuals with observed survival times, some of the observations may be censored and there may also be more than one individual who fails at the same observed time[7]. We suppose that there are g ($g \leq n$) failure times amongst the individuals, and arrange these times in ascending order into. $0 < t(1) < t(2) < t(3) < t(4) \ldots\ldots\ldots < t(r)$

Within each interval, calculate probability of dying within that interval e.g. Interval 4 is (t(3), t(4)] - includes t(4) but not t(3)

Probability of dying in interval 4 is(number of deaths in interval 4) /(number alive at time t(3))

So probability of surviving beyond interval 4 = S(t(4))

S(t(4)) = probability of surviving beyond interval 3 x probability of surviving interval 4

S(4) = S(3) x (1- probability of dying in interval 4)

Recursive relationship - S (1) =1

We count the total number of individuals alive at the start of the interval (ni, i = 1, 2... g) and the number of individuals who died (d_i) in the time interval. The Kaplan-Meier estimate of the survivor function is given by

$$\hat{S}(t_g) = \prod_{i=1}^{g} \left(\frac{ni - di}{ni}\right) \quad \ldots (4)$$

With the approximate standard error (Greenwood's formula)

$$s.e.\{\hat{S}(t_g)\} = [\hat{S}(t_g)]\{\sum \frac{d_i}{n_i(n_i - d_i)}\}^{0.5} \quad \ldots (5)$$

Once $\hat{s}(t)$ is estimated, we can estimate the median survival time $\hat{t}_M$ such that $\hat{s}(\hat{t}_M) = 0.5$ .for different groups of individuals such as male and female, we can estimate a survival function for each group and plot them for comparison. The hazard rate is estimated as

$$\hat{h}(t) = \frac{d_i}{n_i(t_{i+1} - t_i)} \quad \ldots . \quad (6)$$

Cox Proportional Hazard Model

A Cox model is a well-recognized statistical technique for exploring the relationship between the survival of a patient and several explanatory variables. Survival analysis is concerned with studying the time between entry to a study and a subsequent event (such as death)[1]. Censored survival times occur if the event of interest does not occur for a patient during the study period. A Cox model provides an estimate of the treatment

Effect on survival after adjustment for other explanatory variables. It allows us to estimate the hazard (or risk) of death,

The regression method introduced by Cox as [2]:

$$h(t) = h_0(t).\exp(b_k X_k)$$

$$h(t) = [h_0(t)]\, e^{(b_1X_1 + b_2X_2 + \ldots b_kX_k)} \quad \ldots (7)$$

The quantity h0 (t) is the baseline or underlying hazard function, and corresponds to the probability of dying (or reaching an vent) when all the explanatory variables are zero in a Cox model an arbitrary [3]. The baseline hazard function is analogous to the intercept in ordinary regression. The regression coefficients (B's) give the proportional change that can be expected in the hazard, related to changes in the explanatory variables [4]. They are estimated by a complex statistical method called partial maximum likelihood. Assumes that changes in levels of the independent variables will produce proportionate changes in the hazard function, independent of time, or constant relationship between the dependent variable and the explanatory variables is called proportional hazards. If we divide both sides by $h_0(t)$, we get equation(8) which shows where the term proportional comes from.

$$\frac{h(t)}{h_0(t)} = e^{(b_1X_1 + b_2X_2 + \ldots b_kX_k)} \quad \ldots (8)$$

If the hazard ratio = 1 then the variable does not effect on survival [8].

If the hazard ratio < 1 then the variable is associated with

Increased survival.

If the hazard ratio is > 1 then the variable is associated with Decreased survival [9].

The equation 8 implies that the i individual's survival function is a constant power of the baseline survival function.

$$S_i(t) = [S_0(t)]^{e^{(b_1X_1 + b_2X_2 + \ldots b_kX_k)}} \quad (9)$$

Log-Relative Hazard

$$\ln\left(\frac{h(t)}{h_0(t)}\right) = b_1X_1 + b_2X_2 + \ldots + b_kX_k$$

A positive regression coefficient for an explanatory variable means that the hazard is higher and thus the prognosis worse. Conversely, a negative regression coefficient implies a better prognosis for patients with higher values of that variable [9].

2- Practical part

2-1Data description

1- Recording the data of patients infected with hepatitis (B) disease from Gastrointestinal and Hepatic diseases teaching hospital for the years (2002 - 2005) , the number of patients are (100) .

2- The explanatory variables corresponding to each patient are

X_1 denoted the age of the patient.

X_2 denoted the gender of the patient (1= male, 2 = female).

X_3 Occupation of the patient (1=working, 0= not working)

X_4 survival times (days) of the patients at the hospital .

X_5 status a dummy variable indicating whether a case is terminal or censored (1 = death, 0 = censored).

2-2Results and Discussion

1-from applied of Kaplan- Meier estimate of the survival function, Table 1 shows the survival times arranged in sending order (column A). The number of patients who are entering the study is 100 (column B). Since two patients die at the first day (column C). the probability of dying by first day is 2/100 =.0204(column E).So the corresponding probability of surviving up to first day is 1 minus the probability of dying (1-.0204=.9796)(column F).Some survival times are censored(column D). Cumulative probability of surviving up to two day is the probability of surviving at two day and surviving Throughout all the preceding time intervals ie,(.9796x.9548= .9353) (column G). This is the Kaplan—Meier estimate of the survivor function. Sometimes there are censored survival times which occur at the same time as deaths. The censored survival time is then taken to occur immediately after the death time when calculating the survivor function. A plot of the Kaplan—Meier estimate of the survivor function (Figure 1) is a step function, of males and females curves. Figure (1) show that the maiden survival time (days at hospital) for males is 10 days and percentage of staying 75%, for females is 7 days and percentage of staying60%. The comparison of two survival distribution using log-rang test show that highly significant different them.

Statistics	(degree of freedom) df		significant
Log – rank	16.79	1	.0000

Figure(2)show that the mortality of male patients reach a maximum rate in day 11 ,in which the risk was 95%, for female patients it was 50% , but in day 12.

2- T- test[2] was applied between mean age of patients have work and patients with out work and it show that highly significant different 2- groups.

	Mean age	Num.	T-test	df	sig.
Patients working	53	43	2.26	98	0.02
Patients not working	46	57			

3- The first feature to note in table (2), is the sign of the regression coefficients. A positive sign of age means that the hazard (risk or death) is higher, for patients with higher values of that variable. Estimated hazard or risk of death increases by exp(0.026)=1.026 with one year of age adjustment for the effects of the other variables in the modle.for the second variable (gender) the risk of death increases by exp(.357)=1.428 if the patients is male because of the males is more infected than females .for third variable (occupation)risk increases by exp(.245)=1.27 if the patient is working.

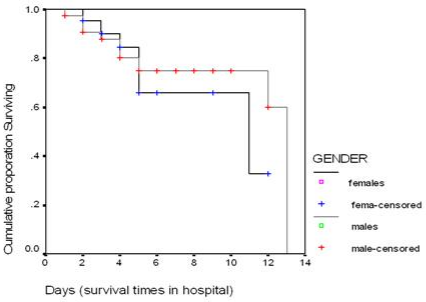
Table (1) calculation of Kaplan – Meier estimate of the survival function.

A Time (days)	B Number of patients at start of study	C Number of deat of the patient h	D Number of censored	E Probability of death	F Proportion survival	G Cumulative Proportion survival
1	100	2	4	.0204	.9796	.9796
2	94	4	11	0452	.9548	.9353
3	79	3	11	.0408	.9592	.8971
4	65	5	19	.0901	.9099	.8163
5	41	4	9	.1096	.8904	.7269
6	28	2	5	.0784	.9216	.6698
7	21	1	3	.0513	.9487	.6355
8	17	1	2	.0625	.9375	.5958
9	14	1	4	.0833	.9167	.5461
10	9	1	1	.1176	.8824	.4819
11	7	1	0	.1429	.8571	.4130
12	6	1	3	.2222	.7778	.3213
13	2	2	0	1.0000	.0000	.0000

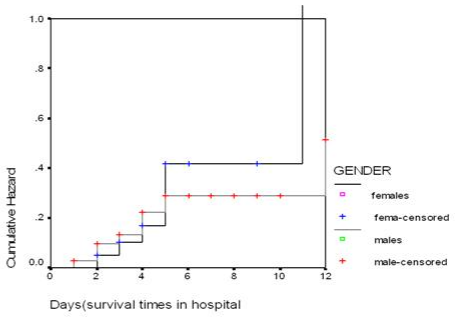
The median survival time for these data is 10.72.

Table (2) Cox regression model and relative risk (exp (B)).

Variables in the Equation	B	SE	Wald	Sig.	df	Exp(B)
Age	.026	.013	4.261	.039	1	1.026
Gender	.357	.521	.469	.023	1	1.428
Occupation	- 245	.439	.310	.017	1	1.277



Figure(1) Kaplan – Meier survival curve .



Figure(2) Kaplan – Meier estimates of hazard function .

References

1.Cox, D.R. 1972.Regression Models and life table. , J. R.Stat.soc. B.34 (2): 187-220.

2.Collett, D.1999. Modeling Survival Data in Medical Research. Chapman & Hall, London, second edition pp.55-58.

3.John Fox ,2002 .Cox proportional – Hazard Regression for survival data appendix to an R and S-plus companion to Applied Regression 1:1-8.
4.Bhattacharjee,A. Lin,D.Y.1994.Cox regression analysis of multivariate failure time data. The marginal approach. In Statistics Medicine 13:2233-2247.
5.Mario, C.2006.Ordered departures from proportionality. Computational sati and Data Analysis 47:517-536.
6.Bhattacharjee, A. 2004.Estimation in hazard regression models under departures from proportionality, Journal of infection diseases, America, 191:182-192.
7.Therneau, T. & Grambsch, P. 2000. Modeling Survival Data: extending the Cox model. Springer 4(2):2234-2237.
8.Singer and John B. 2008 .Cox proportional Hazard Model. SPSS Textbook Examples: Applied Longitudinal Data Analysis London 3(4):1265-1267.
9.Bhattacharjee, A.2007. A Simple Test for the Absence of Covariate Dependence in Hazard Regression Models .University of St. Andrews, 1283-1314.

تقدير نموذج المخاطرة النسبية لبيانات البقاء لمرضى اللتهاب الكبد الفايروسي في العراق

فضاء عثمان سمير*

*وحدة الابحاث البايولوجية للمناطق الحارة /كلية العلوم/ جامعة بغداد

الخلاصة

استخدمت طريقة (Cox Regression model) لتقدير نموذج المخاطرة النسبية وذلك لايجاد العلاقة بين المتغيرات التفسيرية المترافقة للمريض ووقت البقاء .وقد شملت هذه المتغيرات التفسيرية (العمر , جنس المريض ,اوقات البقاء للمرضى ,الحالة النهائية للمريض) . تم تسجيل المرضى المصابين باللتهاب الكبد الفايروسي من مستشفى امراض الكبد والجهاز الهضمي في بغداد للسنوات(2002-2005) .تم تطبيق طريقة (Kaplan-Meier) لتقدير دالة البقاء ودالة المخاطرة .