UNIT 17 COX PROPORTIONAL HAZARD MODEL

Structure

17.1 INTRODUCTION

In the previous units of the survival analysis, we discussed the fundamental concepts of survival analysis as well as methods for analysing and summarising survival data, such as:

- concept of survival analysis and censoring,
- definition of survival, cumulative distribution, probability density and hazard functions,
- estimation of survival functions for complete as well as censored data,
- construction of the Kaplan-Meier survival curves for different groups
- log-rank test for comparing two or more survival distributions/functions.

The Kaplan and Meier survival curve with summary statistics and the log-rank test are powerful and simple methods to compare two or more survival curves, either when they are coming from different treatment arms or different socio-demographic groups like gender, region, etc. These methods are good for the univariate analysis of any set of survival data.

However, in many situations, our problem will be more than this. We may have two groups to compare in terms of their survival experience but they are otherwise also different in terms of other basic characteristics like gender, disease severity at the beginning, age at entry into the study, etc.

Also, in biostatistics, we are generally concerned with the study of the relationship between an outcome (dependent or exposure) variable and one or more independent variables. For example, suppose we want to study how the height and weight of a group of individuals are associated. As height increases, does weight also increase? If so, what is the functional form of such a relation? To site another situation: is gender and weight associated? A third situation could be: does gender and treatment outcome associated? You may note that

the type of statistical method used to study this type of relationship will depend on the type of outcome (dependent) and independent variables.

In the situations discussed above, if the dependent and independent variables are continuous, we usually use correlation and linear regression to study such a relationship. These are described in Blocks 2 and 3 of MST-002: Descriptive Statistics, respectively. This is the case when we study the association between the height and weight of a group of individuals because both are continuous. We consider height as the independent variable and weight as the dependent variable. Here, we measure the correlation (association) between these two variables by the Pearson's correlation coefficient. The correlation coefficient gives a measure of how much these two variables are linearly related.

Now to answer the second question, given in the first paragraph, i.e., as height increases, does weight also increase? We require the functional form of the relation between weight (dependent variable) and height (independent variable). For that, we obtain linear regression and the linear relationship between height and weight can be as

$$\text{Weight} = \text{Constant} + \beta \times \text{Height}$$

where β is called the regression coefficient and can be estimated from the given data using the method of estimation (curve fitting) which has already been described in Unit 5 of the MST-002 course. The relationship shown above will help us to predict weight for a given height.

Let us consider another situation where the dependent variable is again weight but more than one continuous independent variable, say, height and age are considered. That is, there are two independent variables and one dependent variable and all three variables are continuous. Here, we prefer to do a multiple regression analysis. We have also learnt multiple regression in detail in Block 3 of the course MST-002. In this case, linear regression or the linear relationship between age, height and weight can be as

$$\text{Weight} = \text{Constant} + \beta_1 \times \text{Height} + \beta_2 \times \text{Age}$$

where β_1 and β_2 are regression coefficients corresponding to the height and age respectively. For different values of height and age, we can predict the weight.

The above concept can be generalised for any number of continuous variables and we can fit a multiple regression. The general form of relation will be

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

There are other situations where we want to study the relationship between two categorical variables, say, gender and disease prevalence. Usually, gender, which has two categories (binary variable) namely males and females, takes as the independent variable. The disease status which is measured as disease present or absent is again a binary variable and is taken as the dependent variable. Here both are categorical and we use a chi-square statistic for assessing the association between them. In the case of the binary or ordinal outcome variable, we use logistic regression to get the functional form of the relation. This relation is established in terms of probability. The logistic regression also you have studied in Unit 10 of this course.

Here, suppose our dependent variable is the same disease status, however, we have many independent variables, namely gender classified as males and females, socio-economic status classified as low, middle and high, family history of the disease classified as present or absent, age in years, etc. What we mean is we have one dependent variable as binary or categorical and many

independent variables as categorical, continuous and discrete. Here, we go for a multiple logistic regression model to study the association and predict the probability of having the disease. Using the logistic model, we are able to derive the probability of having each state of the dependent variable for different options of the independent variables. The multiple logistic regression you have studied in Unit 12 of this course.

In all above situations, the random variable dealing with the outcome was a complete one or at least assumed as complete. However, there are other situations where the random variables are not complete always or we will end up with incomplete observations. For example, consider a follow-up study where the objective is to study the time to recurrence of myocardial infarction (MI). Here, the study subjects are the people who have already had a myocardial infarction and they are followed up to see the recurrence of myocardial infarction. The information is typically obtained from a follow-up study. Suppose the study has been designed for 15 years follow up. At the end of 15 years follow up, there will be the people/patients for whom the event happened and there will be the people/patients for whom the event did not happen. There will be many individuals (subjects) who do not experience the outcome or event and for such subjects the time variable or dependent variable is incomplete. Thus, for the dependent variable, we are likely to end up with an incomplete data set with many independent variables like age, gender, history of the previous attack, etc. Our earlier explained methods of multiple regression and logistic regression do not find applicable for such data.

Consider another situation where our interest is to study the effect of a new treatment in comparison with an existing treatment. Here, the treatment effect is measured in terms of cured or not and how much time it is taken to cure, duration of the cure. This outcome variable is likely to have incomplete observations for subjects who did not get cured. As earlier, independent variables could be categorical and continuous. Here again, the usual logistic regression does not work and it is an ideal situation for survival analysis. An alternative method for dealing with such situations is the Cox proportional hazard model, which works for both quantitative predictor variables and categorical variables.

In this unit, we will discuss the methodology involved in this method. In Sec. 17.2, we describe the situations for which we can not use regression and logistic regression and why we require the Cox Proboeton hazard model. In Sec. 17.3, we give a brief description of the Cox proportion hazard model. In Sec. 17.4, we focus on the hazard ratio and its interpretation. The estimation and testing procedure of the Cox regression coefficients are explained in Secs. 17.5 and 17.6, respectively. We take an example to show how we apply the model in Sec. 17.7. In Sec. 17.8, we describe 17.8 concept of competing risk. A brief summary of what we have covered in this unit is described in Sec. 17.9. If you will face some problems in the solution of the exercises then you can go through Sec. 17.10 in which we provide the solutions/answers to the exercises. Like the previous unit, our discussion will more concentrate on working with examples than theoretical.

Objectives

After reading this unit, you should be able to

- describe the basis of the Cox proportional hazard model;
- explain the logic of construction of the Cox proportional hazard model;

- provide the description of the data required for the model;
- estimate the Cox regression coefficients;
- test the significance of the Cox regression coefficients; and
- interpret results of the Cox proportional hazard model for a set of data.

17.2 PROBLEM DESCRIPTION

We initiate our discussion with an oral cancer study where the investigators thought of studying the effect of two different types of treatment for oral cancer. Under the first treatment, **Treatment 1**, modality patients received only surgery and we denote this as a **Surgery group**. Under the second treatment, **Treatment 2**, modality initially, the patients underwent surgery and then received chemotherapy for a fixed period of time. Let us denote this as **Chemotherapy group**. The study is designed as a randomized controlled trial. After initial screening for inclusion and exclusion criteria, the subjects were randomized to surgery group or chemotherapy group based on the appropriate randomization method. This is a hypothetical study and the simulated data is given in Table 1.

Table 1: Survival data of 32 Oral cancer patients who underwent two types of treatments

Patient No.	Gender (0 =Male, 1=Female)	Age (in years)	Comorbidity (0 =No, 1=Yes)	Treatment (1 = Surg., 2 = Surg.+Ch.)	Months of Follow up	Event (0 = Death, 1 = Censored)
1	0	60	0	1	8	1
2	1	50	0	1	8	1
3	1	43	1	1	12	1
4	0	62	1	1	12	1
5	1	59	0	1	14	1
6	1	61	1	1	14	0
7	1	71	0	1	15	0
8	0	45	1	1	16	1
9	0	45	0	1	16	0
10	1	51	0	1	17	0
11	0	56	1	1	18	1
12	0	60	1	1	20	0
13	1	58	1	1	22	0
14	0	60	1	1	15	1
15	1	50	0	1	17	0
16	1	45	0	1	24	1
17	0	42	0	2	12	0
18	1	65	1	2	14	1
19	0	58	1	2	15	1
20	0	69	1	2	16	0
21	0	55	1	2	18	0
22	0	55	0	2	19	1
23	0	63	0	2	20	0
24	1	67	0	2	22	0
25	0	49	0	2	22	0
26	1	52	1	2	18	0
27	0	45	0	2	20	0
28	1	42	1	2	23	0
29	1	60	1	2	24	1
30	1	40	1	2	24	0
31	1	49	1	2	24	1
32	1	50	0	2	24	0

There were 32 oral cancer patients randomized into two groups: surgery group and chemotherapy group. The duration of follow up was measured in months. The variable 'event' represents censoring status. Event '0' means death happened and event '1' means censored. The Kaplan and Meier survival curve with the log-rank test is adequate, in case the data had only the variables namely treatment, months of follow up and event and our objective is to compare the effect of treatments. The methodologies of these analyses were well described in Units 15 and 16.

However, in the present data, we have a few more variables measured namely gender, age and other existing morbidity conditions of these patients. Firstly, there may be a difference in the distribution of these variables in both groups. Secondly, these variables are also suspected to have a role in the outcome. Hence, we call them **covariates**. We will first have a description of the covariates in group-wise.

The quantities that potentially affect patient prognosis are known as *covariates.*

Table 2: Summary table of the covariates

Covariates		Treatment 1	Treatment 2
Age	Mean	54.75	53.81
	SD	8.42	9.25
Gender	Male	7 (44%)	8 (50%)
	Female	9 (56%)	8 (50%)
Comorbidity	Yes	8 (50%)	7 (44%)
	No	8 (50%)	9 (56%)

The above-described summary table shows that on average Treatment 2 (surgery + chemotherapy) has younger patients in comparison to Treatment 1 (surgery). Similarly, Treatment 2 has more males compared to Treatment 1. This discrepancy in covariates might have an impact on the outcome. In addition to it, different groups are getting different treatments and outcomes are likely to change accordingly also. That means, we observe a difference in the survival experience between two groups it could be the contribution of difference in treatment and discrepancy in covariates. Then the immediate question will be **how to deal with such data?**

As we discussed in the introduction, in the case of a continuous outcome variable, we would have gone ahead with multiple regression analysis and for a categorical outcome, we would have done a logistic regression to evaluate the effect of treatment after adjusting for covariates (explanatory variable, independent variable, regressor). But here, our outcome variable is incomplete and has two components namely the continuous part which deals with time to the event and a categorical part which deals with event status. Hence, the methods like multiple regression and logistic regression fail to work.

Sir David Roxbee Cox in 1972 suggested the Cox proportional hazard model which is one of the methods for analyzing such type of data. It is a well-recognized statistical technique for exploring the relationship between the survival of a subject/patient and several explanatory variables. This is precisely a regression model in which the predictor, explanatory and independent variables/ factors are usually termed as **covariates** in the survival-analysis literature.

17.3 DESCRIPTION OF COX PROPORTION HAZARD MODEL

Before describing the proportional hazards model, we make a quick revisit to the hazard function which we had defined in Unit 14.

The term **"hazard"** refers to the probability that the subject/patient, under observation at time t, has an event at that time. The hazard function on the other hand is the conditional failure rate. The hazard function is the conditional failure rate of a subject/patient in a small interval after time t conditional to it has survived up to t.

To illustrate this concept, consider a subject whose survival time is 5 years. This means to die in the 5th year the subject has to be alive up to 4 years. That means hazard at the 5th year is conditional that the subject lived up to the 4th year. This concept is incorporated into the computation of the hazard function.

The hazard function is a conditional probability dealing with the instantaneous failure rate. This is also known as **age-specific failure rate** or **conditional mortality rate**. This function is not really a probability since its value can be greater than one. This hazard function plays a very important role in the analysis of survival data and the application of the method of the Cox regression model.

The Cox regression model extends survival analysis methods to assess simultaneously the effect of several risk factors on survival time.

Now, our interest is to explore the possibility of defining a regression like structure to hazard function which incorporates the covariates to study the time to event with censoring information. There are various approaches different people proposed and the most important and simple one applied in health science is the one proposed by Sir David Roxbee Cox.

A Cox proportional hazards model is a well-recognized statistical technique for exploring the relationship between the survival of a subject/patient and several covariates. It evaluates simultaneously the effect of several covariates on survival. The principle of the Cox proportional hazards model is to link the survival time of a subject/patient to covariates. For example, in the medical domain, we are seeking to find out which covariate has the most important impact on the survival time of a patient. In other words, it allows us to examine how specified covariates influence the rate of a particular event happening (e.g., infection, death) at a particular point in time. This rate is commonly referred to as the hazard rate. In our case, we will be able to compare the survival experience of two treatments after adjusting for gender, age and comorbidity. In a proportional hazards model, the unique effect of a unit increase in a covariate is multiplicative with respect to the hazard rate. These models could describe a situation such as a drug that reduces a subject's immediate risk of having a stroke, but where there is no reduction in the hazard rate after one year for subjects who do not have a stroke in the first year of analysis.

Sir David Roxbee Cox
(15 July 1924 -18 Jan. 2022)
British Statistician and Educator

The regression model was proposed by Sir David Roxbee Cox (1972) for investigating the relationship between survival time and the covariates. According to Cox, the hazard function is defined as the product of an arbitrary non-negative baseline hazard $h_0(t)$ and an exponential linear function of the covariates $X_1, X_2, ..., X_p$. The form of the model is given as follows

$$h(t \mid X) = h_0(t) \exp\left(\beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p\right) \quad \ldots(1)$$

The hazard function that is proposed by Cox has two factors:

- $h_0(t)$ - is a non-negative function of time t and does not involve the covariates. It is called the baseline hazard function and can be regarded as the hazard function for a subject/patient whose covariates all have values of 0. In other words, it characterizes how the hazard function changes as a function of survival time in the absence of covariates.
- $\exp\left(\beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p\right)$ - is a function of the set of covariates $X : X_1, X_2, ..., X_p$ and does not involve time. The effect of the covariates on the hazard is modelled as a linear function in this equation which is then exponentiated. The coefficients β_1, β_2, …, β_p are called Cox regression coefficients.

This model is known in the literature by various names like the **Cox model**, the **Cox proportional hazard model** or simply the **proportional hazard model**.

If we take the log of both sides of the Cox model given in equation (1), then we get

$$\log h(t \mid X) = \log h_0(t) + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

In the general case, we think of the ith subject/patient having a set of covariates $X_i : X_{1i}, X_{2i}, ..., X_{pi}$ and we model their hazard rate as some multiple of the baseline hazard rate:

$$h_i(t \mid X_i) = h_0(t) \exp\left(\beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi}\right)$$

This linear function is also known as the prognostic index for the ith subject/patient which is exponential to insure a non-negative hazard.

The Cox proportional hazards model is often called **"semi-parametric"** because it makes a parametric assumption concerning the effect of the covariates on the hazard function but makes no assumption regarding the nature of the hazard function itself. The Cox model assumes that covariates act multiplicatively on the hazard function but does not assume that the hazard function is exponential model, normal or any other particular form.

Here $X : X_1, X_2, ..., X_p$ is a set of covariates of interest, which may include:

- continuous scale factors (age, blood pressure),
- binary nominal scale factors (gender, marital status),
- possible interactions (age by sex interaction)

The coefficients β_1, β_2, …, β_p are called Cox regression coefficients and are estimated from the data. If the sign of the Cox regression coefficients is positive (negative) then its means that the hazard (risk of death) is higher (lower). If any particular coefficient βi is zero, then it implies that the associated covariate X_i is not related to survival. To test whether a particular covariate is related to survival, we test the null hypothesis that the corresponding regression coefficient β (or set of regression coefficients) is equal to zero. If the null hypothesis is rejected, then the data provide evidence that this particular covariate is associated.

17.3.1 Assumptions

The Cox proportion hazard model is based on two main assumptions

1. The ratio of the hazard rates of two subjects/patients corresponding to the two different sets of covariates, known as the hazard ratio, should be

constant over time, or equivalently, that the hazard for one subject/patient/ individual is proportional to the hazard for any other subject/patient/ individual, where the proportionality constant is independent of time. Due to this, the model is called the proportional hazard model.

2. The relationship between the log hazard and each covariate is linear, which can be verified with residual plots.

Now, you may try the following exercises.

E1) Describe the Cox proportional hazard model.

E2) Write the assumptions of the Cox proportional hazard model.

17.4 HAZARD RATIO AND ITS INTERPRETATION

Under the Cox proportional hazard model, we also compute a ratio of the hazard functions, known as hazard ratio (HR). We can define the hazard ratio as the ratio of the hazard rates of two subjects/patients corresponding to the two different sets of covariates. In general, the hazard ratio for covariate at the observed state $X = Y : X_1 = Y_1, X_2 = Y_2, ..., X_p = Y_P$ versus at the observed state $X = Z : X_1 = Z_1, X_2 = Z_2, ..., X_p = Z_P$ is defined as

$$HR\left(t \mid X = Y, X = Z\right) = \frac{h(t \mid X = Y)}{h(t \mid X = Z)}$$

$$= \frac{h_0(t)\exp\left(\beta_1 Y_1 + \beta_2 Y_2 + ... + \beta_p Y_p\right)}{h_0(t)\exp\left(\beta_1 Z_1 + \beta_2 Z_2 + ... + \beta_p Z_p\right)}$$

$$HR\left(t \mid X = Y, X = Z\right) = e^{\beta_1(Y_1 - Z_1) + \beta_2(Y_2 - Z_2) + ... + \beta_1\left(Y_p - Z_p\right)} \quad ...(2)$$

Here, we observe that the hazard ratio is independent of the baseline hazard function and only depends on the function of the set of covariates and regression coefficients of the Cox model. Therefore, it is unimportant what shape the baseline hazard function has. We can compute the hazard ratio using the regression coefficients of the Cox model. It can reflect the comparison of two different sets of covariate values or a single covariate. We use the hazard ratio because it correctly interprets the regression coefficients. It plays the same role in interpreting and explaining the results of the proportional hazards regression model as the **relative risks** and **odds ratio** in logistic regression. But it differs in the sense that the relative risks and odds ratio are cumulative over an entire study, using a defined endpoint, while the hazard ratio represents instantaneous risk over the study time period. The hazard ratio is useful when the risk is not constant with respect to time. It uses information collected at different times. The term is typically used in the context of survival over time.

Suppose there is a single covariate, say, treatment. The treatment can have two options surgery and surgery combined with chemotherapy as in our example. Suppose the first patient is from the surgery and the second from surgery plus chemotherapy and our interest is to assess the hazard function of surgery plus chemotherapy in comparison to surgery. Without loss of generality, we can assume that X = 0 for surgery and X = 1 for surgery plus chemotherapy. Then we can compute the hazard function for both treatments using equation (1) as

$$h(t \mid X = \text{surgery} = 0) = h_0(t)\, e^{\beta_1 \times 0} = h_0(t)$$

Similarly,

$$h(t \mid X = \text{surgery} + \text{chemotherpy} = 1) = h_0(t)\, e^{\beta_1 \times 1} = h_0(t)\, e^{\beta_1}$$

Now, by the definition of the hazard ratio, we calculate it for the patients who had surgery plus chemotherapy in comparison to surgery taking hazard function for surgery plus chemotherapy in the numerator and hazard function for surgery in the denominator as

$$HR\left(t \mid X = 1, 0\right) = \frac{h(t \mid X = 1)}{h(t \mid X = 0)} = \frac{h_0(t)\, e^{\beta_1}}{h_0(t)} = e^{\beta_1}$$

This gives the hazard for the patient who had surgery plus chemotherapy in comparison to the patient who had only surgery provided all other covariates are missing or constant. The hazard ratio is the relative risk and it is interpreted the same way the relative risk is interpreted for a cohort study.

Now suppose in addition to the treatment status we want to include one more covariate namely gender. We are interested in obtaining the hazard ratio for females in comparison to males. This means males are taken as reference (in general equal to 0) and females (in general equal to 1) to be compared with that of males. Hence, the hazard ratio is

$$HR\left(t \mid X_1 = 1, 0, X_2 = 1, 0\right)$$

$$= \frac{h(t \mid X_1 = \text{surgery} + \text{chemotherapy} = 1, X_2 = \text{female} = 1)}{h(t \mid X_1 = \text{surgery} = 0, X_2 = \text{male} = 0)}$$

$$= \frac{h_0(t)\, e^{\beta_1 + \beta_2}}{h_0(t)} = e^{\beta_1 + \beta_2}$$

This is the combined effect of surgery and chemotherapy on females in comparison to surgery alone in males.

So far we have discussed only binary or categorical covariates. However, the Cox model allows us to use continuous covariates. Now let us introduce a continuous covariate. Suppose there is a continuous covariate, say age, in the model then how the hazard ratio works?

In the case of the absence of other covariates, a subject with age 50 versus another subject with age 60, the hazard ratio is

$$HR\left(t \mid X_1 = \text{age} = 60, 50\right) = \frac{h(t \mid X_1 = \text{age} = 60)}{h(t \mid X_1 = \text{age} = 50)} = \frac{h_0(t)\, e^{60\beta_1}}{h_0(t)\, e^{50\beta_1}} = e^{10\beta_1}$$

Now, we try to understand the interpretation of the hazard ratio. In our example, if the Cox regression coefficient when we consider surgery plus chemotherapy in comparison to surgery comes as $\beta_1 = 0.75$ then the hazard ratio will be

$$HR\left(t \mid X = \text{treatment}\right) = e^{-0.75} = 0.47$$

Thus, the hazard ratio = 0.47 means that at a given instant in time the patients with surgery plus chemotherapy have 0.47 chance of dying in comparison to the patients with the treatment surgery only. That is, the patient with surgery plus chemotherapy has less chance of dying in comparison to surgery. Hence, we can say that the surgery plus chemotherapy group has more survival than the surgery only. Another example, suppose the hazard ratio for cancer patients corresponding to the covariate smoking and nonsmoking comes out 3 then this

indicates that the hazard for someone with the risk factor smoking is three times as large as that for those without smoking when all other things are equal.

A hazard ratio greater than one indicates that as the value of the covariate increases, the event hazard increases and thus the length of survival decreases. In other words, if the hazard ratio is > 1, for a treatment group A in comparison to treatment group B then it indicates that treatment group A has a shorter survival than treatment group B, and if it is < 1, it indicates that treatment group A is less likely to have a shorter time to the event than the treatment group B. Hazard ratio of 1 implies equal hazard in the two groups.

Now, you may try the following exercises.

E3) Define hazard ratio.

E4) If a survival study shows the hazard ratio as 2 for a certain type of cancer patients corresponding to the covariate female with respect to male when other covariates are missing then interpret the hazard ratio.

In order to make an inference about the covariates and their effect on time to event or survival time, the whole problem reduces to the estimation of β's. In the next section, we concentrate on the estimation of these coefficients.

17.5 ESTIMATION OF COX REGRESSION COEFFICIENTS

The best way to think about the estimation of parameters in the Cox proportional hazard model is the maximum likelihood estimation. The derivation of the likelihood estimates is beyond the scope of the present section. However, we try to understand the summary of the process.

Suppose we have 'n' independent observations of the time to event with some observations being complete and others incomplete (censored). When we find the maximum likelihood (ML) function for the survival data (complete and censored), it comes as the function of the Cox regression coefficients and baseline hazard function and we have to maximise the same with respect to both. But the form of the baseline hazard function is not specified therefore we cannot use the full maximum likelihood function to estimate the Cox regression coefficients. Cox observed that if the proportional hazards assumption holds (or, is assumed to hold assumption no. 1) then it is possible to estimate the regression coefficients (parameters) without any form of consideration of the hazard function. He proposed a likelihood function called the **partial likelihood function** that does not depend on the baseline hazard function and depends only on the regression coefficients only. The log partial likelihood function proposed by Cox is given as follows:

$$L_p(\beta) = \sum_{i=1}^{m} \left\{ \beta X_{(i)} - \log \left[\sum_{t_{(j)} > t_{(i)}} e^{\beta X_j} \right] \right\}$$

where m - the distinct ordered survival times

$X_{(i)}$ - the value of the covariate for the subject/patient at ordered survival time $t_{(i)}$

$\sum_{t_{(j)} > t_{(i)}}$ - the summation is over all subjects/patients at risk at time $t_{(i)}$.

Once, we have the likelihood function then we can use the standard procedure of getting the estimators by maximizing the likelihood function. That is, we find partial derivatives of the log of the likelihood function $L_p(\beta)$ with respect to each regression coefficient (parameter) in the model, and then solve a system of equations as we have done in Unit 6 of the course MST-004: Statistical Inference. But here we get non-linear equations which are very difficult to solve analytically. Therefore, we use the iteration method. That is, we obtain the solution in a stepwise manner, which starts with a guessed value for the solution, and then successively modifies the guessed value until a solution is finally obtained. These mathematical derivations are beyond the scope of this unit. The interested can get these derivations from Kleinbaum DG (1996) or Hosmer and Lemeshow (2000). From an applied statistics point of view, we concentrate more on the interpretation of the estimates. Once, we obtain the ML estimates of the regression coefficients, then we can obtain the ML estimate of HR using the invariance property of ML estimators, therefore, the ML estimate of the hazard ratio is given as

Invariance property of ML estimation:

If $T = t(X_1, X_2, \ldots, X_n)$ is a ML estimator of θ and $\gamma(\theta)$ is a one to one function of θ, then $\gamma(T)$ is a ML estimator of $\gamma(\theta)$. This is known as invariance property of ML estimator.

$$\hat{HR} = e^{\hat{\beta}}$$

It has been derived that the distribution of β's is asymptotically normal with mean estimated by the maximum partial likelihood estimates of β and variance-covariance matrix estimated by the second derivative of the likelihood function. Once we have the sampling distribution of the estimator, then the standard error can be derived as usual.

17.5.1 Confidence Interval

A confidence interval(CI) is the range of values that is likely to include the true value of the parameter with a certain probability and is used to measure the precision of the estimate. We calculate the confidence interval for the Cox regression coefficients (for a large sample) using the following formula

$$\hat{\beta} \pm Z_{\alpha/2} SE\left(\hat{\beta}\right) \quad \ldots(3)$$

where, $Z_{\alpha/2}$ is the value of standard normal variate at α% level of significance and its value for 99% and 95% confidence intervals are 2.58 and 1.96 respectively.

Now, we can obtain $(1 - \alpha)$% confidence interval for the hazard ratio as

$$e^{\hat{\beta} \pm Z_{\alpha/2} SE\left(\hat{\beta}\right)} \quad \ldots(4)$$

A confidence interval is said to be more precise if it is narrower than the other confidence intervals. The interpretation of the confidence interval of the hazard ratio is similar to that of relative risk. If the confidence interval is totally below unity for a covariate then the covariate is significant of low risk and if it is totally above unity then the corresponding covariate is a risk factor. If the confidence interval includes unity, then the corresponding covariate is not significant.

17.6 TESTING OF COX REGRESSION COEFFICIENTS

After estimating the Cox regression coefficients, a question may arise in our mind "How can we check whether a covariate has a significant impact on the hazard of a subject/patient or not?". For that, we use the test of significance of the Cox regression coefficients.

For testing whether the impact of the covariate is significant or not, we follow the procedure of testing the hypothesis. For that, we formulate the null and alternative hypotheses about the Cox regression coefficient as

$$H_0 : \beta_i = 0 \text{ and } H_1 : \beta_i \neq 0$$

For testing the above hypothesis, we use the Wald statistic which is given as

$$Z = \frac{\hat{\beta}_i}{SE\left(\hat{\beta}_i\right)} \quad \ldots(5)$$

Under some assumptions, the Wald statistic follows a standard normal distribution.

Some software uses another test statistic which is the square of the Wald statistic which is given as

$$\chi^2 = Z^2 = \left[\frac{\hat{\beta}_i}{SE\left(\hat{\beta}_i\right)}\right]^2 \quad \ldots(6)$$

This statistic follows the chi-square distribution with (k – 1) degree of freedom if a covariate has k levels.

We take the decision about the null hypothesis as we have taken in the course MST-004. That is if the calculated value of the statistic is greater than the tabulated value, that is, lies in the critical region, at a given level of significance, we reject the null hypothesis at that level of significance otherwise we do not reject the same. Similarly, we can use the p-value for the same and if the p-value is less than the given level of significance (α), we reject the null hypothesis.

Most of the statistical packages have a routine survival analysis and provide direct estimates of β's and standard error of β's.

17.7 APPLICATION

Now let us call back the data given in Table 1. We use R software to analyse the data using the Cox proportional hazard model. The results are shown in the following table:

Covariates	Estimate of β	SE of estimate
Gender -Female v/s Male	– 0.656	0.647
Age	0.036	0.038
Comorbidity- Yes v/s No	0.336	0.592
Treatment- 2 v/s 1	–1.367	0.601

The above table gives the basic estimates for the Cox regression coefficients in the model and their standard error (SE). Now based on these estimates we will compute other inferential statistics.

The covariates considered for this analysis are given in the first column. Of these, gender, comorbidity and type of treatment are categorical covariates whereas age was kept as a continuous variable. In this analysis, males, no comorbidity and treatment surgery were taken as the reference category and correspondingly hazard ratio was obtained for females, surgery + chemotherapy and presence of comorbidity. We have already defined the estimate of hazard ratio as e^{β}. Using this expression, we can compute the hazard ratio for all covariates as.

Covariates	Estimate of β	HR = e^{β}
Gender -Female v/s Male	– 0.656	0.519
Age	0.036	1.037
Comorbidity- Yes v/s No	0.336	1.399
Treatment- 2 v/s 1	–1.367	0.255

We will try to interpret the hazard ratio. In gender, males are taken as the reference category and the hazard ratio (0.519) for that is less than unity. This indicates that females are at less risk compared to males. Similarly, the hazard ratio for Treatment 2: surgery plus chemotherapy in comparison to Treatment 1: surgery is 0.255 which is also less than 1 and indicate that the patient with surgery and chemotherapy are at less risk compared to only surgery. In other words, we can say that surgery plus chemotherapy is a better treatment than only surgery. The hazard ratio for age and comorbidity is greater than the unity. Therefore, the presence of comorbidity in comparison and larger age have a negative impact on the survival of the patients.

We do not stop our inference just by looking at the hazard ratio. In the next step, we will go for the 95% confidence interval for the hazard ratio. We already have a standard error of the regression coefficients. Moreover, it is known that for a large sample, the regression coefficient follows a normal distribution. Hence, we can compute the 95% confidence interval for β. Once we have the 95% confidence interval for β's, we can take exponentially to get 95% confidence interval for the hazard ratio. We compute the CI in the following table.

Covariates	Estimate of β	SE of estimate	HR = e^{β}	95 % CI for HR	
				Lower	Upper
Gender -Female v/s Male	– 0.656	0.647	0.519	0.146	1.844
Age	0.036	0.038	1.037	0.962	1.117
Comorbidity- Yes v/s No	0.336	0.592	1.399	0.439	4.465
Treatment- 2 v/s 1	–1.367	0.601	0.255	0.078	0.828

This table gives the upper and lower limit of 95% confidence interval of the hazard ratio. The interpretation of the confidence interval of the hazard ratio is similar to that of relative risk. If the 95% confidence interval contains 1, then we interpret that the corresponding covariate is not significant otherwise the covariate is significant. In our case, only the CI of the treatment covariate is less than the unit therefore it shows that treatment is the only significant covariate. Other covariates are not significant.

Now we move to the final part of the inference namely the test of significance. We can also test which covariate has a significant impact on the survival of the patients using the test of significance. We formulate null and alternative hypotheses for the Cox regression coefficients as

$$H_0 : \beta_i = 0 \text{ and } H_1 : \beta_i \neq 0$$

For testing the null hypothesis, we compute the value of the test statistics (chi-square), critical (tabulated) values at 1 degree of freedom at 5% level of significance and p values in the following table.

Covariates	Estimate of β	SE of estimate	Test statistic	Critical value	P value	Inference
Gender -Female v/s Male	– 0.656	0.647	1.028	3.84	0.235	Inignificant
Age	0.036	0.038	0.898	3.84	0.269	Insignificant
Comorbidity- Yes v/s No	0.336	0.592	0.322	3.84	0.598	Insignificant
Treatment- 2 v/s 1	–1.367	0.601	5.174	3.84	0.013	Significant

Since the p-value (0.013) for treatment is less than 5% level of significance, therefore, the treatment covariate is the only significant covariate.

For better understanding, you may try the following exercises.

E5) An analysis is conducted to investigate differences in all-cause mortality between men and women participating in the certain heart study. A total of 158 participants aged 50 years and older are followed until the time of death or up to 12 years, whichever comes first. The participants have different covariates: age, sex, systolic blood pressure, smoking, total serum cholesterol and diabetes. Our goal is to determine which covariate influences the survival time. After applying the Cox regression hazard model analyses, we get the following results:

Risk Factor	Parameter Estimate	SE
Age	0.117	0.006
Sex -Males v/s female	0.403	0.106
Systolic Blood Pressure	0.017	0.003
Smoking-Yes v/s No	0.768	0.104
Total Serum Cholesterol	– 0.002	0.002
Diabetes- Yes v/s No	– 0.203	0.144

i) Obtain hazard ratio and interpret the results.

ii) Find the 95% confidence interval for the hazard ratio.

iii) Test whether the covariates are significant or not at 5% level of significance.

17.8 COMPETING RISK

In survival analysis, we include censored observations where subjects/patients are lost to follow up or the study period ends before the event of interest has occurred and the subjects/patients are expected to experience only one type of event during follow-up, such as death from a brain tumour. If we want to estimate the brain tumour death rate over time then we can use the Kaplan-Meier method and if we want to know whether the mortality rate of brain tumour patients differs between two or more treatment groups then we can use the log-rank method. And if we are interested to find the impact of the covariates on the survival time of the patients then we use Cox proportional hazard model. In these methods/approaches, we assume that the censoring is noninformative. In other words, the survival time of a subject/patient is assumed to be independent of a mechanism that would cause the patient to be censored.

Oftentimes, a patient may experience an event other than the one of interest which alters the probability of experiencing the event of interest and

subjects/patients may be exposed to multiple types of a single event. For example, in the case of brain tumour follow-up study if the death of the patients is our event of interest, then our observations-possibly die from the brain tumour, or heart attack or COVID-19 or even an accident. When only one of these different types of events can occur then such types of events are called **"competing events"**, in a sense that they compete with each other to bring the event of interest, and the occurrence of one type of event prevents the occurrence of the others.

The events which prevent the occurrence or modify the risk of the primary event or outcome of interest are called competing events.

The probability of competing events is known as **"competing risk"**, in the sense that the probability of each competing event is influenced by the probability of other competing events. However, a patient experiencing a competing event is censored in an informative manner. Hence, we may not use traditional methods such as Kaplan-Meier because they are not designed for multiple causes to the same event. For handling competing events/risks we use techniques such as cumulative incidence function (CIF), etc. which are beyond the scope of this course.

We end this unit by giving a summary of its contents.

17.9 SUMMARY

1. A Cox proportional hazard model is defined as

$$h(t \mid X) = h_0(t)\exp\left(\beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p\right)$$

   The Cox proportional hazard model is also known **Cox model**, the **Cox proportional hazard model** or simply **proportional hazard model**.

2. The ratio of the hazard rates of two subjects/patients corresponding to the two different sets of covariates is known as hazard ratio. If

   HR = 1: No effect

   HR < 1: Reduction in the hazard

   HR > 1: Increase in hazard

3. The log partial likelihood function proposed by Cox is given as

$$L_p(\beta) = \sum_{i=1}^{m}\left\{\beta X_{(i)} - \log\left[\sum_{t_{(j)} > t_{(i)}} e^{\beta X_j}\right]\right\}$$

4. The (1–α)% confidence interval for the regression coefficients is given as

$$\hat{\beta} \pm Z_{\alpha/2} SE\left(\hat{\beta}\right)$$

5. The (1–α)% confidence interval for the hazard ratio is the exponential of the CI for the Cox regression coefficients and is given as

$$e^{\hat{\beta} \pm Z_{\alpha/2} SE\left(\hat{\beta}\right)}$$

6. For testing the hypothesis regarding the Cox regression coefficients, we use Wald statistic which is given as

$$Z = \frac{\hat{\beta}}{SE\left(\hat{\beta}\right)} \text{ or } \chi^2 = Z^2 = \left[\frac{\hat{\beta}}{SE\left(\hat{\beta}\right)}\right]^2$$

7. Estimation procedure for parameters
8. Interpretation of proportional hazard model computation output.
9. Competing events and competing riak

17.10 SOLUTIONS/ANSWERS

E1) Refer to Section 17.3.

E2) Refer to Sub-Section 17.3.1.

E3) Refer to Section 17.4

E4) The HR = 2 means that the male patients have two times the chance of dying in comparison to the female patients with cancer.

E5) As we know, the estimate of hazard ratio is e^{β}. Therefore, we can compute the hazard ratio for all covariates/risk factors as

Risk Factor	Parameter Estimate	HR = e^{β}
Age	0.117	1.124
Sex -Males v/s female	0.403	1.497
Systolic Blood Pressure	0.017	1.017
Smoking-Yes v/s No	0.768	2.155
Total Serum Cholesterol	– 0.002	0.998
Diabetes- Yes v/s No	– 0.203	0.816

Since the hazard ratio for age risk factor is 1.124 so we can conclude that the expected hazard is 1.124 times higher in a person who is one year older than another holding other covariates/risk factors as constant. Similarly, the hazard ratio = 1.497 for males in comparison to females indicates that the males have 1.497 times higher risk as compared to females, holding other covariates/risk factors as constant. In the same way, the hazard ratio for systolic blood pressure and smoker are 1.017 and 2.155 respectively, therefore, we can say that hazard is 2.155 times higher in a person who is a smoker in comparison to a non-smoker.

Now, we compute the 95% confidence interval for the Cox regression coefficients/ parameters (β). Once, we have the 95% confidence interval for β's, we can take exponentially to get 95% confidence interval for HR. We calculate the same in the following table.

Risk Factor	Parameter Estimate	SE	HR = e^{β}	95 % CI for HR	
				Lower	Upper
Age	0.117	0.006	1.124	1.111	1.137
Sex -Males v/s female	0.403	0.106	1.497	1.215	1.844
Systolic Blood Pressure	0.017	0.003	1.017	1.012	1.022
Smoking-Yes v/s No	0.768	0.104	2.155	1.758	2.642
Total Serum Cholesterol	–0.002	0.002	0.998	0.995	1.001
Diabetes- Yes v/s No	–0.203	0.144	0.816	0.615	1.083

This table gives the upper and lower limit of 95% confidence interval of HR. Since CIs for total serum and diabetes contain 1 so these risk factors are not significant. Rest are significant.

Now, we test the significance of the risk factors using the testing of hypothesis. We formulate null and alternative hypotheses for the Cox regression coefficients as

$$H_0 : \beta_i = 0 \text{ and } H_1 : \beta_i \neq 0$$

For testing the same, we calculate the Wald statistics which follow the chi-square distribution with one degree of freedom for all risk factors with their critical values and p-values in the following table.

Risk Factor	Parameter Estimate	SE	Test statistic	Critical value	P value	Inference
Age	0.117	0.006	387.876	3.840	0.000	Significant
Sex -Males v/s female	0.403	0.106	14.355	3.840	0.000	Significant
Systolic Blood Pressure	0.017	0.003	44.941	3.840	0.000	Significant
Smoking-Yes v/s No	0.768	0.104	54.624	3.840	0.000	Significant
Total Serum Cholesterol	–0.002	0.002	1.699	3.840	0.131	Insignificant
Diabetes- Yes v/s No	–0.203	0.144	1.986	3.840	0.105	Insignificant

Since the p-values for age, sex, systolic blood pressure and smoking are less than 5% level of significance, therefore, these risk factors are significant whereas total serum cholesterol and diabetes are not significant.